EMERGENCY MEDICAL SERVICE RESPONSE SYSTEM PERFORMANCE IN AN URBAN SOUTH AFRICAN SETTING: A COMPUTER SIMULATION MODEL

Christopher Owen Alexander Stein

Thesis Presented for the Degree of

DOCTOR OF PHILOSOPHY

In the Division of Emergency Medicine

UNIVERSITY OF CAPE TOWN

March 2014

DEDICATION

For Leonie, Coleen, Ryan and Euan. Words cannot express how much I value your support in allowing me the time to complete this work, and in having put up with my absence for so long.

ACKNOWLEDGEMENTS

To my supervisor, Lee Wallis, for his guidance, advice and support.

To my co-supervisor, Olufemi Adetunji, for his technical advice, guidance and patience with my drawn-out approach to simulation modelling.

To Prof. Sarma Yadavalli, for kindly assisting me in finding a co-supervisor.

To Shaheem De Vries, Dexter Timm and Stanford Nomdo for assistance with access to information, critique and advice on various parts of the simulation model.

To Paul Hansen for extracting, supplying and explaining the incident data used for modelling.

To Simio LLC, for providing me with a grant to use their software.

To Prof. Ahmed Fethi for supplying me with a licenced version of ArcGIS.

To the Faculty of Health Sciences and colleagues from the Department of Emergency Medical Care, for supporting me for the duration of this study, and in particular during my six months of sabbatical leave.

This research was supported in part by the National Research Foundation of South Africa (Unique Grant 86454).

ABSTRACT

This study investigated the effects of different response strategies, vehicle location strategies and vehicle numbers on response times in a simulated Emergency Medical Services system. The simulation was a computer model using discrete-event simulation and the model was based on Western Cape Emergency Medical Services operations in Cape Town. The study objectives were to (i) create the simulation model, (ii) determine the best-performing combination of explanatory factors and (iii) determine the effect of increasing vehicle numbers on response time performance. The simulation model took into account incident arrival rates, incident and hospital spatial distributions, vehicle numbers and dispatch practices in the modelled system. Verification and validation of the simulation model utilised a combination of quantitative and qualitative methods. The validated simulation model was changed in two ways: (i) the response strategy was changed to either single- or two-tier (the response model factor) and (ii) the vehicle location strategy was changed to either dynamic or static (the vehicle location factor). This yielded four individual models each representing one combination of these factors. Each simulation model was run for a simulated period of seven days. Output data were analysed using multivariate analysis of variance in order to identify differences in response time between the factor combinations. A single-tier model using dynamic vehicle locations produced the best response performance. This model was run repeatedly, increasing vehicle numbers incrementally with each run to assess the effect of increased vehicle numbers on response time performance. A doubling of vehicle numbers resulted in an 14% increase in the number of responses meeting the national performance target for high acuity incidents, while a seven-fold increase in vehicle numbers increased this to 15%. No further performance increases were seen beyond this with increased vehicle numbers. A 2% performance increase for lower acuity incidents was seen with the same increase in vehicle numbers. In the system modelled, increasing vehicle numbers should not be expected to realise anything more than small improvements in response time performance, at a high operational cost. Fine-grained dynamic deployment of vehicles in anticipation of system demand appears to be a more important determinant of response performance than vehicle numbers alone.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DEFINITION OF KEY TERMS

Advanced Life Support (ALS)

The level of care provided by emergency care personnel holding the Critical Care Assistant, National Diploma in Emergency Medical Care, Emergency Care Technician or Bachelor's Degree in Emergency Medical Care qualifications. ALS can also refer to a level of equipment resourcing required by emergency care personnel with these qualifications to practice within their scope.

Attribute

"A property of an entity."(1)

Basic Life Support (BLS)

The level of care provided by emergency care personnel holding the Basic Ambulance Attendant qualification. BLS can also refer to a level of equipment resourcing required by emergency care personnel with this qualification to practice within their scope.

Computer Model

"[A] simulation model implemented on a computer."(2)

Conceptual Model

"A non-software specific description of the simulation model that is to be developed, describing the objectives, inputs, outputs, content, assumptions and simplifications of the model."(3) Also sometimes referred to as a simulation model.

Discrete-Event Simulation

"The modelling of systems in which the state variables change only at a discrete set of points in time", [linked to events in the system].(1,4) "The system state can change at only a countable number of points in time…at which an event occurs, where an event is defined as an instantaneous occurrence that may change the state of the system."(5)

Dynamic [Vehicle Location Factor Level]

Referring to Emergency Service Vehicles (ESVs) positioned at decentralised holding points based on proximity to high incident density or demand. Available ESVs are also moved to cover holding points where all ESVs are unavailable.

Emergency Medical Care (EMC)

"Patient care for an acute condition; this care can be given by a variety of different medical practitioners, including technicians, nurses, paramedics, physician assistants and/or physicians."(6)

Emergency Medical Services (EMS)

"The system that organizes all aspects of care provided to patients in the prehospital or out-of-hospital environment. The term 'EMS' in context may encompass or refer to local, regional, national or international systems for delivery of patient care."(6)

EMS System

"[A] System that provides for the arrangement of personnel, facilities and equipment for the effective and coordinated delivery in an appropriate geographical area of health care services under emergency conditions."(7) Either EMS or EMS System can be used to refer to EMC provision at the systems level, depending on the context.

Emergency Service Vehicle (ESV)

Any vehicle used as part of a response system, to transport patients to hospital and/or respond to incidents. These vehicles are dedicated to EMS use and custom built and equipped for this purpose. In this study two types of ESV are referred to. Ambulances can respond to an incident and transport one or two lying patients to hospital. Primary Response Vehicles (PRVs) can respond to an incident but are not configured for transportation of any patients. Ambulances can be staffed and equipped to either ALS or ILS/BLS level of care, while PRVs are typically staffed and equipped only to ALS level of care.

Entity

"An object of interest in [a simulated] system".(1)

Exempt

An incident not requiring transportation of any patients to a hospital, either because no patient was located at the incident scene or because there was no requirement for transportation (the patient may have refused transportation or may have been declared dead at the incident scene, as examples).

Incident

An event associated with one or more patients experiencing some kind of condition requiring EMC.

Intermediate Life Support (ILS)

The level of care provided by emergency care personnel holding the Ambulance Emergency Assistant qualification. ILS can also refer to a level of equipment resourcing required by emergency care personnel with this qualification to practice within their scope.

Model

"A representation of a system for the purpose of studying that system."(1)

Model Validation

"The process of determining whether a simulation model is an accurate representation of the system, for the particular objectives of the study."(8)

Model Verification

"Concerned with determining whether the conceptual simulation model has been correctly translated into a computer [model]."(8) "…to assure that the conceptual model is reflected accurately in the [computer] model."(9)

Out-of-Hospital Care

"Medical care provided to patients who are located in settings other than a hospital; generally this applies to patients who are not planned or intended to be transported to a hospital."(6)

Pre-hospital Care

"Medical care provided to patients in settings other than a hospital and who are planned or intended to be transported to a hospital for further care or evaluation."(6)

Primary Response Vehicle (PRV)

A non-transport ESV used in two-tier response models. PRVs are typically fewer in number than ambulances, cover larger areas and are staffed by ALS-level practitioners. PRVs may or may not travel with an ambulance to hospital from an incident scene, depending on whether the patient(s) being transported require ALS-level care during transport.

Priority 1 Incident (P1)

An incident that, according to information provided by an individual reporting it, is of such a nature that it requires the most urgent response and ALS level of care. P1 incidents are generally associated with patients experiencing an acutely limb- or life-threatening condition. There may be an actual or imminent threat to the airway and often severe abnormalities of vital homeostatic functions such as pulmonary gas exchange, fluid, electrolyte and acid-base balance or cardiac rhythm and tissue perfusion. The descriptor "P1" can also be used to refer to the state of an individual patient.

Priority 2 Incident (P2)

An incident that, according to information provided by an individual reporting it, is of such a nature that it requires a less urgent response. P2 incidents are generally associated with patients experiencing a non-life-threatening or even minor condition, and who are currently stable and able to compensate physiologically, with or without treatment such as supplemental oxygen, intravenous fluids or other medication, including intravenous analgesia. The descriptor "P2" can also be used to refer to the state of an individual patient.

Process

"A time-ordered sequence of interrelated events separated by intervals of time, which describes the entire experience of an entity as it flows through a system."(10)

Quality

"Quality of care is the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge."(11)

Response System

That part of EMS concerned with ensuring a co-ordinated response to a reported incident. Includes ESV management, call taking, dispatch, ESV response to the incident scene and transportation of one or more patients to hospital.

Response Time

The time interval spanning the beginning of call taking (i.e. answering a call for EMS assistance and recording the relevant details) until the arrival of an ESV at the associated incident location.

Simulation

"The imitation of the operation of a real-world process or system over time".(1)

Single-tier [Response Model Factor Level]

Referring to a system utilising only transport ESVs (i.e. ambulances) which are generally a mix of ALS and non-ALS capability in some ratio.

Static [Vehicle Location Factor Level]

Referring to ESVs positioned at fixed, centralised locations (relative to a sector) not explicitly chosen for their proximity to incident density or demand.

System

"A group of objects that are joined together in some regular interaction or interdependence toward the accomplishment of some purpose".(1)

Two-tier [Response Model Factor Level]

Referring to a system utilising both transport ESVs (ambulances) and non-transport ambulances (PRVs) in some ratio. Typically, PRVs are far fewer in number than ambulances, cover larger areas, are associated with ALS capability and are allocated only to P1 incidents. Ambulances are generally a mix of ALS and non-ALS capability in some ratio.

Vehicle Availability

The proportion of ESVs available for allocation to an incident at a given time in a particular EMS, or part thereof.

CHAPTER 1: BACKGROUND

1.1. INTRODUCTION

This study is centred on the investigation of factors affecting response system performance in a large, urban Emergency Medical Services (EMS) system in South Africa. Response performance relates to how rapidly an EMS system can respond to a request for Emergency Medical Care (EMC) by processing the request, allocating an appropriate Emergency Service Vehicle (ESV) and having that ESV travel to the scene of an incident.

In this Chapter, the rationale behind a rapid EMS response to requests for EMC will be presented by arguing that timeliness is a fundamental component of quality EMS provision, and that optimising response performance should therefore be a priority in any EMS system. This is followed by a formal definition of the Emergency Response Interval (ERI) and its components. The origins of response time as an EMS performance indictor, and the choice of a threshold value for high acuity response time in North American EMS systems is discussed. Having defined the response interval and established the place of response time as a performance indictor in EMS systems, three factors that may influence response performance in urban South African EMS systems, and that are key variables in this study, are introduced.

The background information summarised above culminates in a statement of the research problem, which is centred on a lack of information concerning the functioning of EMS response systems in South Africa, and how system factors may modulate response performance in order to achieve national response time benchmarks. This is followed by a concise statement of the research question, the study's aim and objectives and an overview of the methodology. Because this study uses computer simulation rather than data gathered from observations in a real EMS system, this approach is defended and finally delimitations of the study are presented before an outline of the thesis structure concludes the Chapter.

1.2. TIMELINESS AS AN ATTRIBUTE OF QUALITY EMERGENCY MEDICAL CARE

EMC is defined as any form of patient care for an acute condition delivered by a broad spectrum of health care professionals in and outside of hospitals.(6) EMC occurring outside of the hospital environment may be in the form of out-of-hospital care, where the primary intention is the provision of EMC which may be followed by transfer to a hospital, or pre-hospital care, where the intention is transfer to a hospital for further evaluation and care after initial management at the emergency

scene.(6) It is mainly within the context of the latter that the term Emergency Medical Services (EMS) is used, with reference to systems that organise all aspects of care in this environment.(6)

Usage of the term EMS as suggested above is very broad, referring simply to all aspects of organised care. Closer examination of the literature on EMS systems reveals frequent emphasis on components or attributes of EMS systems as a way of giving a finer-grained notion of what EMS is. These range from communications, transportation and trained personnel to system finance, information systems and research.(6,12,13) More of a process-orientated view of EMS is conveyed by the US National Highway Traffic Safety Administration's representation of functions of an EMS system, which includes incident recognition, EMS access, dispatch and varied levels of response from first responders through to Advanced Life Support (ALS).(14) Although this approach gives an impression of function as well as structure, what EMS "does" and not only what EMS "is", it still does not convey any information about "how" EMS functions, or how EMS might be expected to function from the patient's perspective.

The "how" of EMC delivery by EMS is another way of referring to quality of patient care. Although its importance in EMC seems intuitively obvious, quality is notoriously difficult to define in a meaningful and measurable way. The US Institute of Medicine defines quality in two main ways; by referring to health services bringing about measureable change to improve health outcomes and by referring to health services being "consistent with current professional knowledge".(11) The Institute of Medicine further sets out six quality aims encapsulating very broadly what quality health care "should be" in the future. One of these is that health care should be timely, "*…reducing waits and sometimes harmful delays for both those who receive and those who give care*."(11) In the context of EMS systems, timeliness refers to performance of the response system.

It is on this attribute of quality health care that the current study is focused, within the context of pre-hospital emergency care and specifically EMS system design and functionality. The remainder of this Chapter will be devoted to expanding more fully upon the emergency response and how it is characterised and measured. The link between emergency response time intervals and EMS performance indicators as measures of quality will be explored in more depth following which the problem statement and its significance, research aim and objectives and a broad overview of the study methodology will be given. The Chapter will conclude with a description of the study's delimitations.

1.2.1. The Emergency Response and Emergency Response Interval

Emergency response is a fundamental element of an EMS system(13,14) and typically follows recognition of an incident, public access to the system and some kind of dispatch process. In the broadest sense, “response” can mean many things – ranging from activation of a first responder system, an ambulance, a Primary Response Vehicle (PRV) or even an aircraft. The time interval spent on response as a whole has been broken up into a variety of discrete sub-intervals which are generally a reflection of local system meaning and usage.

Studies aimed at creating and using a model for evaluating response system performance in the US some two decades ago still provide the most comprehensive definitions for the emergency response interval and its sub-intervals.(15,16) According to this model, the total time interval reflecting EMS response activities contains 10 discrete time intervals from notification of an incident through to what is termed the “recovery” interval (activities at the hospital in preparation for return to service).(16) For the purpose of this study, which is focused upon assessment of the effects of several EMS system design choices on the timeliness of emergency care as an indicator of EMS quality, response interval definitions have been based on the model proposed by Spaite, Valenzuela and Meislin.(16) The model has been simplified, and not all of the 10 suggested time intervals have been used. This modified Emergency Response Interval (ERI) model, including events defining each interval and ESV availability, is shown in Fig 1.1.

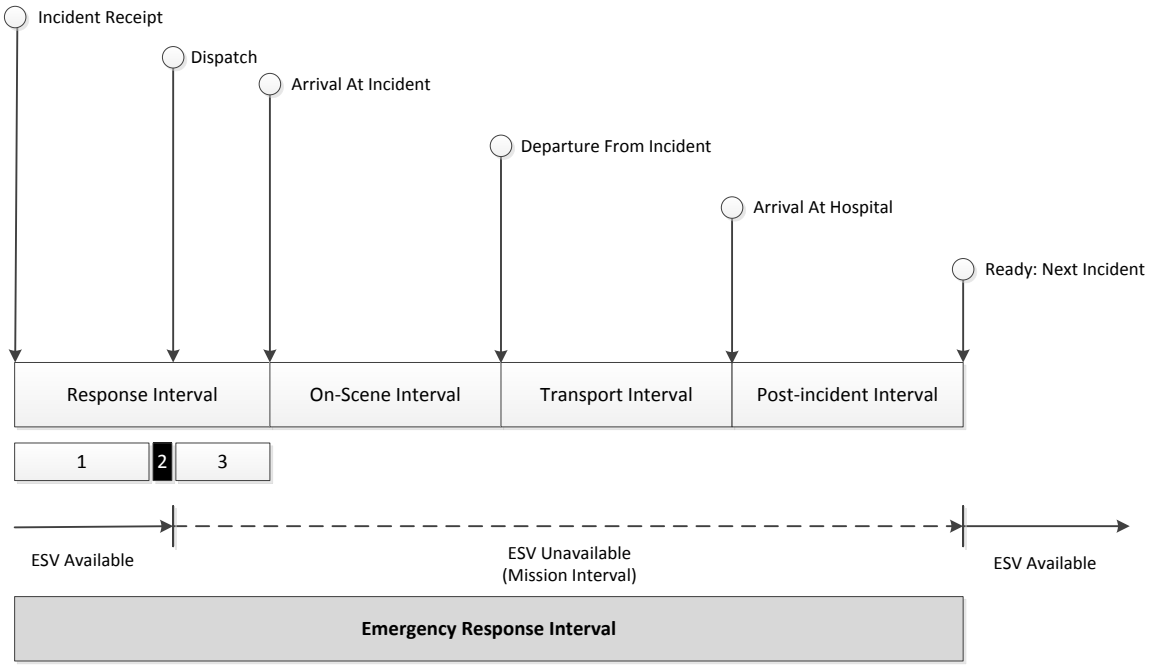


Figure 1.1. Components of the Emergency Response Interval
ESV = Emergency Service Vehicle, 1 = Dispatch Interval, 2 = Hand-off Delay, 3 = Response Travel Interval
Adapted from Spaite, Valenzuela and Meislin.(16)

1.2.2. Emergency Response Time Intervals as Performance Indicators

Evidence first emerging in the late 1970s of the time-sensitive nature of successful defibrillation in cases of sudden cardiac death played a key role in highlighting the importance of EMS access to patients in the shortest possible time.(17) The importance of rapid response to cardiac arrest cases seemed to shape perceptions about response times in general (response time here meaning the time measured for the response interval in Fig 1.1), especially in the USA, where the eight minute defibrillation and four minute cardiopulmonary resuscitation thresholds identified above by Eisenberg, Berger and Hallstrom (17) became the *de facto* response time benchmark for all high acuity pre-hospital cases.(18) The notion of the "Golden Hour" in trauma care added further weight to the idea that the performance of EMS systems, many of which deal with a significant case load of injured patients, should be judged by compliance with response time standards.(19)

Despite the counterintuitive nature of setting response time benchmarks as if EMS response takes only cardiac arrest cases into account, response time still occupies prime position as an indicator of EMS performance in many EMS systems.(20–23) Although criticism has been levelled at the lack of evidence validating the use of response times as an EMS performance indicator, much of this is aimed at the eight minute threshold as mentioned above, rather than the idea that the measurement of response times in general has no value as a performance indictor.(22,24–28) The more important question seems to be how EMS response systems can better differentiate between those conditions where a rapid response does have some impact on final outcome (other than cardiac arrest) and those where it does not, and to tailor the response to the condition. Until evidence exists to guide decision-makers in this regard, response time in its current form and according to its current performance goals will most likely remain a key EMS performance indicator.

1.2.3. Factors Influencing Emergency Response Intervals

A range of factors may potentially influence any of the intervals contained in the ERI. These include, but are not limited to, ESV numbers in the system, type of response system used (single- or two-tier), how and where ESVs are stationed in the response area, whether and how ESV location is changed to anticipate changes in demand, terrain and road conditions, weather conditions, traffic congestion, proximity of hospitals to incident locations, waiting times at hospitals and more.

Investigation of the effect of any of these on any part of the ERI may be of interest in a given system. However some of these factors may be more amenable to change than others, or may be influenced by other system variables that could change them. Three basic system-related factors that could

apply to many urban EMS systems, and that could have an impact on response times, have been chosen as the focus for this study which aims to assess their effects on response times.

1.2.3.1. Number of Emergency Service Vehicles

ESV availability is dynamic in any EMS system, and can be considered to be relative to ESV demand among other factors. As demand for ESVs escalates, an EMS system will be able to accommodate this demand until all ESVs are unavailable at which point emergency calls will be queued until one or more ESVs become available again. In this simplistic "supply and demand" view, ESV availability is a function of ESV numbers and consequently adding more ESVs to a system with low ESV availability is expected to improve response times by decreasing the number of queued incidents.

1.2.3.2. Response Model

Response model refers to whether only ambulances are used for response (single-tier system) or a mix of ambulances and PRVs (two-tier system). In South Africa, because of a long-standing shortage of ALS-qualified personnel, many systems have been structured around a two-tier system with the belief that the best use can be made of a small cohort of ALS-qualified personnel by placing them in PRVs which do not need to be tied up with transportation in every case and thus increase their availability for dispatch to appropriately triaged incidents. Consequently, the South African meaning of two-tier is ALS dispatched first (in a PRV) and Intermediate Life Support (ILS) or Basic Life Support (BLS) dispatched subsequently (in an ambulance) to high acuity incidents, which is quite different compared to the usual meaning of two-tier in North American EMS systems (BLS first, followed by ALS and no PRVs to high acuity cases).(29)

Although the two-tier model as described above is intended to minimise response times and make maximum use of ALS personnel, it is dependent on an adequate number of ambulances and on optimal positioning of all ESVs in order to minimise response times and maximise ESV availability, specifically the availability of ALS personnel. A two-tier system that is under-resourced in terms of transport vehicles (i.e. ambulances) and utilises sub-optimal positioning of ESVs may produce the opposite of what is intended by frequently tying up ALS personnel in PRVs waiting at incident scenes for ambulances and by inappropriately dispatching a small number of PRVs over long response distances.

1.2.3.3. Location and Movement of Emergency Service Vehicles

The geographic location of ESVs relative to the location of emergency calls may affect ESV availability and distances that are required to be covered for each emergency response, thus affecting response times. The idea of locating ESVs strategically closer to geographic areas of historically high incident density in order to reduce response times was first expressed as part of what later became known as System Status Management (SSM), which originated in the US in the late 1980s.(30) SSM is a term used to define a wide variety of techniques to optimise the balance of resources and demand in a system and to minimise response times. In its most complex form SSM involves more than just placing ESVs close to areas of historically high demand, however even this approach is not consistently applied in urban South African EMS systems. This is particularly so in systems where Fire Departments act as agents in the provision of EMS - ESVs are typically located at Fire Stations which are not positioned intentionally in any proximate way to areas of high incident density.

The three factors above are basic resource and EMS system design choices which can logically be thought to have an effect on response times in urban South African EMS systems. Exact details of how these factors characterise urban EMS systems in South Africa are not known, however they can be considered as possibly affecting response times, however they are configured. The nature of their effect on response times is worthy of closer consideration, as this information has the potential to facilitate changes in the approach to EMS systems design and resourcing in order to realise the aim of timely EMC.

1.3. PROBLEM STATEMENT

As described in Section 2 of this Chapter, the avoidance of delays in access to patient care may be considered a quality aim of any health care system. From an EMS perspective, this translates to the delivery of timely pre-hospital EMC. Timeliness is difficult to define in an absolute way, and may differ from patient to patient, however local systems typically set goals with regard to response times that are considered to be important system performance indicators. One way to minimise response times is to optimise ESV availability in an EMS system, which in turn may be dependent on other factors relating to the response model and location of ESVs within the response area. An ideal EMS response system should know and deploy the optimal configuration of these and other factors in order to make the most efficient use of available resources.

Currently in South African urban EMS systems, there appears to be little consistency in system design choices with many practices related to the above factors seemingly determined by adherence

to historical precedent. In most cases, little has changed over time in the way these systems have positioned themselves to provide pre-hospital EMC to the populations they serve. Consequently, across urban South Africa, the relatively modest response time benchmarks set at national level are seldom if ever complied with.(31–33)

Decision-makers in the EMS systems domain whose responsibility it is to address this problem do not currently have much in the way of data to guide them. Little is currently known about the effects of the factors described above on system functioning, and specifically about effectiveness of the response system in an urban South African context. While some data are available from research on EMS response systems in other countries, this is often not applicable to local EMS systems and in most cases does not address the fundamental questions about system design and resourcing that could lead to improvement.

The current study was conceived in order to answer some of these fundamental questions about factors influencing the performance of EMS systems in urban South Africa. The questions being asked focus on effects; the effects of different combinations of the factors mentioned above on response times and ESV availability, and the method employed uses computer simulation of an urban South African EMS environment. Computer simulation remains the only feasible way of comparing alternative EMS systems design choices through their effects on ESV availability and response time as the study and manipulation of real systems is impractical, prohibitively costly and unethical. By providing some of these basic answers and allowing EMS decision-makers to factor at least some evidence into their planning and design, it is hoped that this study will contribute to future efforts in delivering on the promise of timely pre-hospital emergency care in urban South Africa.

1.4. RESEARCH QUESTION

What is the optimal configuration of ESV location, response model and ESV numbers required to meet response time performance goals, or provide the best possible response performance, in a simulation model based on a large urban South African EMS system?

1.5. RESEARCH AIM

This study aimed to investigate the effect of ESV location, response model and ESV numbers on response system performance in a computer simulation model of a large urban South African EMS system.

1.6. RESEARCH OBJECTIVES

The research objectives were:

1.6.1. To create a baseline model of an urban EMS system by means of computer simulation.

1.6.2. To alter the baseline model referred to above in order to create four different models representing all possible combinations of two factors, namely (ii) response model and (ii) ESV location. These two factors each had two levels:

(i) *Response model*: Single- and two-tier, referring to a system with only transport ESVs (i.e. ambulances) and a system with a mix of non-transport (i.e. PRV) and transport ESVs respectively. Both single- and two-tier response models utilise a mix of ALS and non-ALS staffed vehicles.

(ii) *ESV location:* Static and dynamic, referring to a system with ESVs positioned at fixed, centralised locations not related to incident density or demand and a system with ESVs positioned at decentralised holding points based on proximity to high incident density or demand respectively. Dynamic location of ESVs also involves movement between holding points based on real-time demand patterns.

(iii) To determine which of the models above produced the shortest response times and the greatest proportion of P1 response times within 15 minutes and P2 response times within 60 minutes with a baseline number of ESVs. Furthermore, to determine whether the identified model met response time and ESV availability performance goals of 15 minutes or less for at least 90% of high acuity (P1) cases, with no other (i.e. P2) case having a response time longer than 60 minutes.

1.6.3. If the identified model in iii) above did not meet the above response time and ESV availability performance goals, to determine the minimum ESV numbers required in that model to meet this goal.

The research objectives described above were amended slightly from the original set of objectives given in the research proposal for this study. The proposed objectives included incident on-scene time as a dependent variable and an on-scene time of ≤ 20 minutes as one of the response system performance goals. However after input data analysis and consideration of the model's complexity, it was decided to model incident on-scene times solely on input data distributions and parameters, thus precluding its inclusion as a dependent variable.

The proposed set of objectives did not specify exactly how increases in ESV numbers as in 1.6.2. (iii) would be applied to the combination of experimental factors and models described in 1.6.2, because the full extent and complexity of these models were not known. The decision to isolate one best-performing model and assess the effect of increased ESV numbers on this model alone was taken once a better understanding of the complexity of the baseline model had been obtained.

Lastly, ESV availability > 0% at all times was initially listed as one of the response system performance goals in the proposed set of objectives. After validation of the baseline model and observation of data from each of the factor models described in 1.6.2 it became clear that high ESV availability was a feature of all of these models and that availability would not reach 0% at any point in any of the experimental replications. The performance goal of ESV availability being > 0% at all times was thus removed from the study objectives described above.

1.7. OVERVIEW OF THE METHODOLOGY

A detailed account of this study's methodology is given in Chapter 3. The following section gives an overview of the methodology as part of the background to this study, in which a computer simulation model was constructed and used to generate data for analysis. The specific type of simulation used was discrete-event simulation which is characterised by the modelling of a system in which change of the system state occurs only at a discrete set of points in time, linked to events. This type of simulation is well suited for the modelling of EMS systems where changes in system state are event-driven (examples of events could be the receipt of an emergency call at a dispatch centre, the dispatch of an ESV or the arrival of an ESV at an incident). Simulation models are generally either based upon a real system, or a hypothesised one that does not yet exist (where the objective of the simulation may be to investigate the functioning of the proposed system). This study used discrete-event simulation to build a model of the Western Cape EMS Cape Town response system following the step-wise approach set out by Banks *et al*.(1):

(i) *Problem formulation*: Behaviour of the system under consideration was described, together with objects and activities falling within the experimental framework.

(ii) *Conceptual modelling:* A high-level, software independent description of the structure and function of the system was created defining the modelling objectives, objects and related attributes, process logic, assumptions, simplifications and input data distributions. The level of detail in the conceptual model was guided by the modelling and study objectives.

(iii) *Data Collection:* Incident and system-related data were obtained from the Western Cape EMS Computer Aided Dispatch (CAD) system and from system experts. These data were used for input modelling, in order to determine key variables and probability distributions to be used in the computer model.

(iv) *Model translation:* A detailed representation of the system was created, based upon the conceptual model and analysed input data. This was done using a commercial simulation software application called Simio (Design Edition, version 6.97, Simio LLC, Pennsylvania, USA).

(v) *Verification and validation:* The computer model referred to above was verified using a range of techniques, meaning that checks were carried to ensure that the conceptual model had been transformed into a computer model with sufficient accuracy. Validation was performed throughout the development phase, in order to ensure that the computer model was accurate enough for the research problem. This was done by comparing outputs of parts of the model, and the whole model when completed, with outputs of the real system.

The single validated model in existence at this point was used as a basis for further development of four separate models reflecting each of the four combinations of explanatory factors described in 1.6.2. Changes made to the initial model only addressed behaviour determined by the explanatory factors while the rest of the model in each case was left in its original state. Following this, the four models were compared in the following steps:

(vi) *Experimentation:* Each of the four computer models was run over a number of replications in order to generate data on response times.

(vii) *Output data analysis:* Output data were analysed statistically in order to identify which of the four models produced the best combination of response time. The identified model was then run for several further replications, each time adding ESVs, until the specified response time performance goals had been met or no further improvement in response performance was observed.

As described briefly above, computer simulation was used in this study in order compare different types of EMS response systems. The basis for this decision rested upon consideration of a wide range of factors, and took into account the principle that computer simulation is not necessarily a valid solution for any type of research or operations-based problem and that its use must be justified. This justification is given below.

1.8. MOTIVATION FOR THE USE OF SIMULATION

Simulation was chosen as an appropriate method in this study for a number of reasons, some related to the practical and ethical obstacles associated with real EMS systems research and some related to the nature of the research problem. These reasons are presented below. The reasons in 1.8.1 can be thought of as reasons why non-simulation research would be difficult or impossible to carry out, thus making a case for the use of simulation. The reasons given in 1.8.2 relate more directly to the characteristics of this research that make the use of simulation a good option.

1.8.1. Practical and Ethical Considerations

Very little published data exist addressing the effect of any aspect of EMS system structure on performance.(34–36) Typically, these studies are retrospective and do not manipulate EMS system structural factors. Even simple questions about the effect of system design characteristics on performance have few quality evidence-based answers derived from research carried out in real EMS systems. The reasons for this are most likely related to the difficulty of performing experiments at system level in a real EMS system:

(i) *Impracticality and expense:* It is generally not possible, for practical reasons or because of the expense involved, to manipulate system-level variables for the purposes of research. For example, investigating the question of optimal ESV numbers in a real EMS system would involve the provision of increased numbers of ESVs and personnel to study the effects of this. Such expenditure for the purposes of research is beyond the resources available to any EMS in South Africa.

(ii) *Difficulty in measuring performance indicators:* Research on the effectiveness of various EMS system structural configurations would require measurement and analysis of a range of performance indicators as dependent variables. In the discussion above just two performance indicators are suggested: response time and ESV availability. Of these, the first (as defined from the time of call centre receipt to arrival of the ESV on scene) would most

likely present a challenge and the other would probably be very difficult to measure accurately in many EMS system in South Africa in an objective way.

(iii) *Ethics and public perception:* In attempting to manipulate EMS system structural factors for the purposes of studying their effects, it is likely that system performance would be adversely affected at some point. In fact, comparing levels of system performance across varying combinations of structural configuration (some of them almost certain to produce poor performance) may be the only way of differentiating between poor and superior choices in this regard. Manipulating a real EMS in this way would present difficult questions regarding informed consent for users of the system whose health may be affected, and would most likely be considered unethical anyway from a risk: benefit analysis point of view. It could also lead to a public outcry if poor system performance and harm (whether real or perceived) is linked to experimentation.

Given the limitations above, it is clear why virtually everything that is known about system-level behaviour and effects in EMS has been derived from simulation studies.

1.8.2. Characteristics of the Research Problem Making Computer Simulation Feasible

Computer simulation is not the only modelling approach that can be used to answer questions about the operation or performance of a system. Other analytical methods, from paper calculation to spreadsheet modelling and mathematical programming, can be used.(4) However the range of application of these approaches is quite small as they are complex to understand, cannot account for variability and its effects without a significant increase in complexity and are typically associated with very restrictive assumptions.(1,4,37) The complexity of the system modelled in this study and the need to incorporate the effects of variability into the model and its output, along with the need to easily understand details of the model, made simulation the obvious choice.

One serious restriction when considering computer simulation is the absence of system data to use as model input. When this is so, and when there are not even estimates of key data, then attempting simulation is not recommended.(1) In the current study, it was possible to obtain a substantial amount of data from the Western Cape EMS CAD system upon which to base the model. Although not all data were available in this form, it was possible in most other cases to use estimates of time intervals, proportions or other parameters. The availability of data thus made simulation a viable proposition.

The ability to compare different or new designs, patterns and policies without having to change or perhaps even gain access to real systems is cited as an advantage of simulation.(1,4) Comparison of different systems was a focus of the current study, as set out in the objectives, and made the use of simulation an ideal approach. Even if real systems could have been found with the different combinations of factors described in 1.6.2, the restrictions discussed above under 1.8.1 would still apply. In addition, different data collection policies, systems and approaches in those systems would inevitably have created a lack of homogeneity with regard to input data.

For the reasons given above, this study would not have been feasible to carry out in a real EMS system. Although this would seem to leave no other option but to approach the research problem using simulation, some other approaches have been described and might have been used. However considering the nature and complexity of the research problem and the study objectives, simulation emerged as the best choice.

1.9. DELIMITATIONS

Delimitations of a study serve to limit its scope and clearly define its boundaries. Unlike a study's limitations, delimitations are chosen and intentionally stated by the researcher in order to limit the range and the depth of the study to make it feasible, coherent and relevant. Delimitations of this study as they apply to a number of specific areas are presented below.

1.9.1. Objectives

The research aim and objectives are set out in 1.5 and 1.6 above. These are narrowly focused on the comparison of four different types of response system arising from combinations of two different response system design factors, with the focus on response time as an outcome measure. Although research on EMS systems, and even more narrowly on response systems, could include many other objectives, these were chosen because of the emphasis placed nationally on response time as an EMS system performance indicator.

1.9.2. Context of the Aim and Objectives

The context of the research aim and objectives is urban EMS operations. While investigation of the performance of response systems in urban, peri-urban and rural environments is important, it could be argued that urban systems pose the greatest challenge in meeting response time performance goals because of the large, densely populated areas that they serve. In addition, the complexity of urban response systems arises from a different set of circumstances compared to that of rural

systems meaning that research objectives can only realistically be focused on one of these in a given study, particularly when using simulation as a method of investigation.

1.9.3. Response Variables

Response time was chosen as a response variable because it is currently the only nationally defined EMS system performance indicator. Although there is much debate about the clinical significance of response time as a performance indicator (as discussed in the literature review), it remains the preeminent indicator at the present time with clearly defined performance goals specified at national level in South Africa. For this reason, response time and the proportion of responses meeting response time performance targets were chosen as response variables for this study.

1.9.4. Explanatory Factors

A multitude of EMS system and other factors can be theoretically linked to changes in response time and ESV availability. Of all of these, the number of ESVs in a system is perhaps the most basic determinant of availability of those vehicles and consequently of response times. Two other factors have the potential to affect ESV availability and response times, namely the location of ESVs and the response model used (as described in 1.6.2). These three factors were chosen as explanatory variables because of their wide applicability in urban EMS systems in South Africa and because there is currently a lack of knowledge of their effects (either singly, or by interaction) on response times. To a greater or lesser extent, these are three factors that could possibly be changed to bring about improvements in ESV availability and response times if evidence to support their effectiveness in this regard is detected.

It is important to note that one of the most important drivers in EMS systems configuration and resourcing, namely cost, has not been included in this study as a response variable. This was done so as to limit the complexity of the simulation model, within the constraints of the scope of the study. Although this does not negate the value of the model in identifying the effects of the factors discussed above on either of the response variables, implications arising from the study may be constrained in their application by the consideration of cost implications.

1.9.5. Generalizability

By its nature, simulation involves the construction of a model based upon some kind of real system or in some cases, a hypothetical system that does not yet exist. In either case, output data obtained

from the simulation can only really provide information about the functioning of the modelled system taking into account all of the model's assumptions and simplifications.

In the current study, a combination of approaches was used as described in the methodology overview above. Although it is possible to argue that many of the response system processes are fairly generic and could apply to several real systems, the spatial distribution of incident locations and the relationships between incident locations and hospitals in all of the models were based solely upon the modelled real system. Thus it is not possible to generalise response time results from this study directly to any other real system because these spatial relationships will not hold.

1.10. SUMMARY AND OUTLINE OF CHAPTERS

In this Chapter the notion of timeliness as an EMS system performance indicator has been put forward, together with a model indicating the components of the ERI, the most important of which in the context of this study is the response interval. Several factors affecting the duration of the response interval in EMS systems have been described; specifically ESV location, response model and ESV numbers. A lack of information about the effect of these factors on response system performance in South Africa, and the impact that this has on EMS system design, was identified as the essence of the research problem driving this study. The remainder of the Chapter was devoted to a detailed description of the study's aim, objectives and delimitations, and an overview of the methodology. Because this methodology involves the use of computer simulation, a motivation for this was also given.

This thesis is presented in six Chapters. A brief overview of the remaining Chapter content and structure is given below.

Chapter 2 is a review of the literature related to two major concepts underpinning this study. The first is that of timeliness as an EMS performance indicator. The ideas introduced in Chapter 1 are expanded upon and the argument is put forward that, although available data linking response time to improved clinical outcomes is equivocal, response time (particularly in high acuity cases) is still considered to be an important EMS performance indicator at the present time. The second major concept included in the literature review is that of computer simulation, the methodology upon which this study is based. The review covers both the fundamentals of discrete-event simulation and the modelling process, and its applications in the study of response system performance problems in EMS.

Chapter 3 presents the study's methodology. This includes detailed descriptions of a conceptual model derived from the Western Cape EMS Cape Town system, discussion of the approach used for input data modelling and an explanation of how the model translation process (from conceptual to computer model) was executed. This is followed by a description of the model verification and validation procedures used and the data analysis methods employed.

Chapter 4 sets out results derived from the four simulation model outputs, each one representing a combination of the two experimental factor levels (ESV location and response model). Descriptive data from each of these models is followed by results of hypothesis tests assessing the effects of each model on response time and proportional response time target compliance. Based on these results, a best-performing model is identified. Response performance results from the addition of increased ESV numbers to the best-performing model are presented, along with explanatory and other data on ESV availability, the hand-off delay, ALS presence at incidents, response distances and mission times.

Chapter 5 is a discussion of the study's results, their implications and their possible application to real-world EMS system performance in South Africa. The importance of ESV location in determining response system performance is emphasised, and the limited role of increased ESV numbers is clarified.

Chapter 6 is a conclusion, followed by recommendations which relate to possible areas of improvement in EMS system processes and response performance. The recommendations are followed by suggestions for future research.

CHAPTER 2: LITERATURE REVIEW

2.1. INTRODUCTION

In the opening Chapter of this thesis, the importance of timeliness as a feature of quality in health care systems was put forward. This was contextualised by reference to the ERI, response time as an EMS performance indicator and some of the factors that may influence response system performance.

In the first part of this Chapter, response time as an EMS performance indicator is discussed in greater depth, and more critically in order to gain an understanding of whether the time taken to respond to a high acuity incident has some clinical significance. Although, as argued later in the Chapter, there is currently no definitive answer about whether shorter response time "makes a difference" to patient outcomes, the assertion is made that currently the role of response time as an EMS performance indicator is well established and used as a standard against which the quality of such systems is judged, either alone or in combination with other performance indicators. Concrete examples of this use of response time are given, in the form of response time targets adopted in the US, UK, parts of Europe, Australia and South Africa.

The second part of this Chapter is devoted to a description of discrete-event simulation and modelling - the method used for investigation of system factor effects on response performance in this study. After a definition and description of the discrete-event approach to simulation, the important concept of variability in simulation modelling is discussed with emphasis on the use of random numbers. This is followed by descriptions of each step in the discrete-event modelling process, from the formulation of a conceptual model to the important tasks of model verification and validation.

In the final part of this Chapter, published studies utilising simulation as a method for the investigation of EMS-related problems are reviewed. The emphasis is on studies that identified mean response time, or a coverage area for provision of a specified response time, as dependent variables. In order to place the current study in context, this part of the literature review is arranged under sub-headings each of which describe how demand and dispatch, travel and other process were modelled. Approaches used for verification and validation in these studies are also described along with a concluding discussion on the factors that other investigators have found to improve response time.

2.2. QUALITY IN HEALTH CARE AND EMERGENCY MEDICAL SERVICES

Quality in a general health care context is not easily defined, and this is also true of quality in EMS. As summarised by Kallsen and Stroh,(38) various attempts have been made to give an accurate, concise and operationally relevant definition of health care quality however none of these definitions appear to be satisfactory. Apart from the debate about whether quality measurement should focus on outcomes or processes, definitions proposed in the literature that are convincing in their range and depth tend to be hopelessly immeasurable. Equally, those definitions emphasising simplicity and ease of measurement tend to be criticised for being overly simplistic. Perhaps the most widely agreed-upon definition of quality in health care comes from a 1998 statement from the US National Roundtable on Health Care Quality:(39)

> *"Quality of care is the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge."*

This definition, while broad and simple enough to be applied to and understood in any health care setting, does not give an indication of specific areas of health care services that should be considered as particularly important in determining quality of care. In recognising that the US health care system was failing to deliver the benefits that it could, the Institute of Medicine published a report in 2001 detailing the root causes of this problem and setting out a detailed action plan for change. In its opening pages, this report contains a broad statement about the adoption by all health care organisations and professionals of a shared purpose to decrease burden of illness, injury and disability and to improve the health of the US population. In many ways this is the purpose statement of the definition of quality given above, however, much like the definition it lacks focus.

In order to clarify exactly how the Institute's purpose statement could be translated into a more practical "agenda" six performance characteristics were suggested that, if improved, could bring about a realisation of the broader purpose. These six "aims for improvement" are that (quality) health care should be:

(i) *Safe* – that in attempting to help patients they are not injured or otherwise harmed.
(ii) *Effective* – that services provided should be based upon scientific knowledge and that these services are provided to all that may benefit from them.
(iii) *Patient-centred* – that patient values, preferences and needs guide clinical decisions.

(iv) *Timely* – that harmful waits and delays are eliminated.

(v) *Efficient* – that waste of equipment, supplies and energy are avoided.

(vi) *Equitable* – that the quality of care does not depend upon a patient's personal characteristics.

Although the above list does not satisfy the requirements of a general definition of quality, it is a useful summary of the dimensions of health care service where the evaluation of quality might be focused. Equally, it provides some guidance as to where efforts at improving quality may be concentrated.

The bridge between our understanding of what quality is and quality as a demonstrable attribute is the identification of performance criteria, indicators and standards which can be used to make consistent judgements about a system's performance in relation to a given description of quality.(20,40,41) If chosen carefully, defined clearly and validated, these operational indicators of quality can be used as indices of EMS system performance. In simpler terms, quality in EMS today is not something to be guessed about or assumed, but something to be managed. The assertion is made that only the measureable can truly be managed.(20)

2.3. PERFORMANCE INDICATORS AND MEASUREMENT IN EMERGENCY MEDICAL SERVICES

Moore defines a performance indicator as "...*a criterion related to the quality of the program or service that can be measured*."(40) Performance measurement is further defined as "...*quantifiable assessment according to the indicators providing an evaluation and planning tool leading to improvement and quality*." As indicated above, performance indicators can provide evidence of a system's value by providing a continuous measurement of quality, identifying areas of excellence and verifying the effectiveness of corrective action where applicable.(40)

The most commonly cited classification of performance indicators is that originally proposed by Donabedian (adapted here for the EMS setting):(40,41)

a) *Structure:* Performance indicators related to the system's environmental attributes, including characteristics of the physical setting, personnel, equipment, resources and organisational structure. Examples of structure-related performance indicators include quantification and description of available equipment, resources (such as the number of ESVs), system characteristics (such as dispatch model or ESV location) and personnel qualifications. Although

generally the easiest to measure, this class of performance indicators is the most difficult to associate directly with outcomes.(20)

b) *Process:* Performance indicators related to the activities occurring between practitioners and patients during the delivery of emergency care. Moore further defines a process in this context as a "*...repeatable sequence of actions used throughout interrelated components of a prehospital EMS system to produce something of value*."(40) Examples of process-related performance indicators include proportions of patients receiving particular types of treatment or treatment within a specific time frame, and ERI times such as response times, scene times and total pre-hospital time.

c) *Outcome*: Performance indicators related to changes in health status that can be attributed to receiving pre-hospital emergency care. Examples of outcome-related performance indicators include morbidity, mortality and patient satisfaction.

Aside from classification of performance indicators, several other guides exist dealing with their desirable characteristics many of which are fairly extensive. Typical characteristics include relevance to quality of the system, practicality, scientific basis, explicit definition, validity and reliability.(40,41)

2.4. TIMELINESS AS AN ATTRIBUTE OF QUALITY: RESPONSE TIME AS A PERFORMANCE INDICATOR

In defining health care systems quality, the above discussion cited both a broad definition and a more practical listing of performance characteristics put forward by the Institute of Medicine. One of these performance characteristics is that of timeliness; the elimination of waits or delays for both patients and health care professionals. This dimension of quality would appear to be particularly relevant in EMS systems, where at least a subset of cases requiring an EMS response are of a high acuity, time-sensitive nature. Even in less urgent lower acuity cases, the performance characteristic of timeliness would apply in EMS, as it would to non-urgent patients awaiting consultation in a hospital. The perception of timeliness being an important indicator of EMS quality is not limited to the domain of professional health care, it is also prevalent in the lay media and expectations of the public.(42–44)

2.4.1. Response Time Definitions and Standards

The ERI was defined in Chapter 1 (Fig 1.1). Of all the sub-intervals comprising the ERI, four time intervals have emerged as potential candidates for EMS system performance measurement: (i) response time (measuring the response interval), (ii) scene time (measuring the scene interval), (iii) total pre-hospital time (measuring the combined times of all intervals) and (iv) time to appropriate

hospital care (not included in the ERI as shown in Fig 1.1). Of these, response time is most frequently singled-out as representing some form of EMS quality standard.(20,40,41,45)

The start of the response interval is defined in two ways. The first is from the point of view of the practitioner, and begins with the time that the practitioner is mobile to an incident requiring a response.(23) The second is from the point of view of the patient or bystander, and begins with the time the call for emergency assistance is received at a call centre.(41) No one definition predominates and literature incorporating a measure of either of these intervals generally includes an accompanying definition for clarity. In both cases, the end of the response interval is generally taken as arrival at the scene of the emergency.(23,41) Response time is then defined as the elapsed time between the beginning and end of the response interval.

Benchmarks for response time as an EMS performance indicator vary from system to system, but all of these differentiate between urgent cases and non-urgent cases in determining the range of response times considered acceptable. As discussed in Chapter 1, historical factors related to cardiac arrest survival and the emphasis on this form of resuscitation in the early years of EMS shaped much of the thinking about response time benchmarks in North America which appear to persist today.

Benchmarks in the USA specify that an eight minute response time should be achieved in 90% of all high acuity incidents (this is a reference to ALS capability being on scene). This particular definition of response time begins when the responding unit(s) are *en route* to the incident and not when the emergency call is received at a dispatch centre.(23) In the UK the target response time is eight minutes for 90% of category A incidents (high acuity), 19 minutes for 95% of category A incidents and 19 minutes for 95% of category B incidents (lower acuity).(46) Since 2008 response time in the UK has been measured from the time a call for assistance is connected at the dispatch centre (and not when the dispatcher picks up the telephone) until arrival of an EMS vehicle at the emergency scene.(46)

The situation in Europe varies from country to country, however a survey by the European Emergency Data Project (47) found that EMS response time standards varied between four and nine minutes in four European countries.(48) No additional information was given on the precise definition of response time in each case. Australian National standards refer only to how response time as a performance indicator should be measured (using the 50^{th} and 95^{th} percentiles) but do not provide any statement about a compliance standard.(49,50) In South Africa, the response time

performance goal is 15 minutes for 90% of high acuity (Priority 1) cases in urban areas and 40 minutes in rural areas. A performance goal of 60 minutes maximum response time for any case has also been defined. Response time is defined as beginning at the time a call is picked up by a call taker at the dispatch centre until arrival of an ESV at the incident.(31–33) Currently in South Africa, these response time performance goals are the only EMS performance indicators specified at national or provincial level.

2.4.2. Is Response Time a Meaningful Performance Indicator?

Increasingly, authors in the emergency medicine literature have been calling into question the validity of response time as an EMS performance indicator.(22,25–28,51–54) This appears to be partly, and quite logically, because of a growing number of studies suggesting that survival outcomes in a mix of trauma and non-trauma populations are not affected by EMS response times and in particular, whether response times meet a specific performance target or not. A growing emphasis on clinical performance indicators in more recent times may also have contributed to the shift away from response time as an index of EMS system quality.(41) However some counterexamples in the form of evidence suggesting a relationship between response time and patient outcome do exist, making the above trend less certain.(55–57)

The relationship between response time and survival in five mixed patient populations (trauma and non-trauma) has been studied using retrospective research designs. Blackwell and Kaufman included 5,424 patients serviced by an all-ALS EMS agency in a crude analysis of death probability by response time. Their findings were that there was no significant difference between median response times in the survivors vs. non-survivors groups, nor was there a difference between observed and expected deaths. Mortality risk was however significantly reduced for those patients in whom response time was five minutes or less.(24)

Pons *et al.* also retrospectively studied survival to hospital discharge in a sample of 9,559 patients treated and transported to a single Emergency Centre (EC) for further care. Patients were stratified according to mortality risk, based upon assessment of their hospital clinical record. Response times for each incident were obtained and a logistic regression model was used to assess the predictive value of an eight minute response time threshold on patient survival to hospital discharge. A survival benefit was identified in medium- to high-risk cases where the response time was ≤ 4 minutes, but no benefit was evident in these groups when response time was modelled as a continuous variable or when the data were dichotomised on an eight minute response time threshold.(26)

A retrospective cohort design was applied by Blanchard *et al.* to investigate a question similar to that posed by Blackwell *et al.* and Pons *et al.* In this case adult patients with a "life threatening event", as identified at initial triage during the dispatch process, were tracked until hospital discharge or death. A logistic regression model was again used to assess the predictive value of an eight minute response on survival to discharge controlling for patient age, acuity and combined scene and travel time to hospital. No significant survival benefit was detected for a response time ≤ 8 minutes, however some evidence was found of such a benefit for response times of ≤ 7 minutes.(58)

Weiss *et al.* reviewed 2,164 cases of patient transport by a private EMS which fell into four complaint/incident categories: motor vehicle accidents, penetrating trauma, difficulty breathing and chest pain. Response times for each case were analysed, along with vital signs data and the number of vital signs out of range. Outcomes of interest were number of hospital admissions, survival to hospital discharge, number of days in hospital and number of admissions to intensive care. Results showed that response times to trauma cases were significantly shorter than to other complaint categories, however no relationship was found between response time and any of the outcome variables.(28)

Blackwell *et al*. used a case-control design in order to assess whether a locally determined response time target (10 minutes and 59 seconds) had any effect on patient outcomes in a single-tier ALS paramedic service with BLS first responders. Cases (n = 373) were identified as Priority 1 calls (high acuity) with response times ≤ 10:59 while controls were an equally numbered, randomly selected set of calls with longer response times. Outcome measures were mortality and the number of critical interventions performed in the field. No additional mortality risk was identified for Priority 1 patients waiting longer than the specified response time.(25)

The relationship between outcome and response time in trauma cases has also been studied, although not as often. Pons again examined this relationship but this time in a set of 3,576 trauma patients whose data were extracted from a trauma registry.(26) The total set of patients was split by response time (≤ 8 minutes and > 8 minutes) and survival was studied after controlling for age, mechanism of injury and Injury Severity Score. No difference in mortality was found between the two response time groups.(22) A similar result was obtained in a study designed to assess the effect of pre-hospital time on trauma outcomes in urban and rural settings, where response times were

not significantly different in patients who died or survived in the urban grouping. In this study a mortality difference was observed in the rural grouping.(35)

Although the studies discussed above generally conclude that response time has no effect on patient survival in their respective populations, some of them (25,26) have identified beneficial effects, albeit at very short response times and not the performance benchmarks used in the services that were studied. Some other data exist to support the assertion that shorter response time is related to better patient outcomes.(55–57) Although these studies have impressive sample sizes, they are retrospective in nature and thus can only suggest a possible causal link, as is the case for the studies discussed above.

In summary, the available evidence on a causal relationship between response time and patient outcome is equivocal. Although the studies discussed above do not represent the strongest evidence in terms of their designs and methods, it is unlikely that a very different approach would be possible in this area of enquiry. Despite efforts to control response time by minimising it during emergency response, it is not an entity that could be controlled in the manner necessary for rigorous study in the form of a controlled trial and so this quality of evidence will most likely never be accessible in guiding decisions about the clinical meaningfulness of response time as a performance indicator. Because of the notion that timeliness is important in the provision of emergency care, response time will probably be retained for some time to come as a measure of the adequacy of systems delivering this care.

This study takes as a fundamental assumption that response time is a meaningful performance indicator of EMS systems in general, and that investigation and further understanding of the factors affecting response time could improve the quality of service delivery by these systems in general. In motivating for the use of simulation as a way of studying response system performance in EMS systems in Chapter 1, several reasons were given as to why this kind of research would not be possible in real systems and why it is well suited to simulation as a means of enquiry. The rest of this literature review is therefore devoted to providing a background on computer simulation in general and discrete-event simulation specifically, and on how simulation models are constructed, verified and validated. This is followed by a discussion on the applications of simulation to response system problems in EMS that have been reported in the literature.

2.5. COMPUTER SIMULATION IN EMERGENCY MEDICAL SERVICES SYSTEMS RESEARCH

As indicated above, research involving the manipulation of real EMS system variables is generally not feasible. In order to better understand these systems, and to experimentally investigate possible changes to them which may improve performance indicators of interest, a different approach must be taken which does not involve the manipulation or interruption of real-world patient care. Study of an EMS system that does not exist in the real world involves the construction of a system model that produces behaviour accurate enough to be used for a specific set of study objectives. Such a model may be constructed in a number of different ways including the use of mathematical programming approaches, spreadsheet modelling or simulation.(4,5)

The application of analytical solutions (such as mathematical programming or spreadsheet modelling) to the study of complex systems is limited for a variety of reasons. In general, analytical solutions involve the construction of complex models even for relatively simple systems. In addition, although some of the analytical approaches can model system variability, many cannot and those that can do so by assuming less system complexity.(4,5) These reasons, together with the added constraint of comparatively restrictive assumptions for other modelling approaches, make the use of simulation for the modelling of complex systems an attractive choice. Contrary to being historically considered the "method of last resort", simulation today is often considered the only feasible approach due mostly to the sheer complexity systems being modelled.(5) Although simulation may involve physical objects in certain very restricted applications, for the most part simulation is carried out on a computer either by the development of a custom software application or (more commonly) by the use of a commercial simulation software product.(10)

2.5.1. Discrete-event Simulation

Computer simulation models may be static, depicting a system at a particular point in time, or dynamic depicting a system as it changes over time.(4,5) Two further descriptors can be applied to simulation models:

(i) Deterministic or Stochastic

Simulation models containing no random, or stochastic, variables as inputs are considered to be deterministic. The outputs of such models can be determined exactly from the inputs, although this does not necessarily mean that they are less complex than stochastic models. Most complex systems must be modelled to incorporate at least some degree of random inputs and are thus stochastic models. One consequence of utilising a stochastic model is that the model's output will also be

stochastic and will thus represent an estimate of the output variables, necessitating the use of inferential statistical methods for analysis.(1,5)

(ii) Continuous or Discrete

Discrete simulation models are based upon an approach where the model's state changes at a discrete set of points in time. This is in contrast to a continuous simulation model, where the model's state changes continuously over time. These two descriptors can be applied analogously to the systems being modelled, namely discrete systems and continuous systems although as Law and Kelton point out, few systems fit exclusively into one category or the other.(5) Similarly, it is not necessarily the case that discrete systems must be simulated by using discrete simulation models or continuous systems by using continuous simulation models. It is also possible to mix discrete and continuous approaches into a single model. The modelling approach chosen is determined partly by the system characteristics and partly by the objectives of the study driving the modelling process and the nature of output data required for analysis.(1,5)

Discrete-event simulation, the type of simulation approach used for modelling in this study, uses a dynamic systems modelling approach where the model's state variables change only at separate points in time. The points in time at which the model's state changes are those coinciding with an event, which is an occurrence that changes the model's state.(5) As the simulation progresses over time, each event is associated with a model state representing change of the modelled system. Rather than adopting an approach where mathematical methods are used to "solve" for a solution, discrete-event simulation models are run to produce stochastic output data representing an estimate of the true performance of the model.(59)

2.5.1.1. Basic Components of Discrete-event Simulation

The basic functioning of a discrete-event simulation over time can be depicted algorithmically. However this requires the definition of a few fundamental components and concepts required for execution of such an algorithm.(59) Some of these have been defined elsewhere, however these definitions will be given here again briefly.

- **Entity:** An object requiring representation in the model.
- **System State**: A set of variables completely describing the modelled system's configuration at any time.
- **Event**: An occurrence that changes the modelled system's state.

- **Event Notice**: Record of an event that will occur at the current time or some time in the future. The Event Notice is associated with any required data relevant to the event, most importantly the event time.
- **Event List:** A list of Event Notices, ordered by time of occurrence (this is sometimes also referred to as the Future Event List or FEL).
- **Activity:** A time duration, the beginning time of which is known (an Activity is also sometimes referred to as an unconditional wait). Activities may be derived from statistical distributions.
- **Delay:** A time duration, of indefinite length – the Duration length is not known until it ends, often brought about by a set of conditions in the model or by an event (a Delay is sometimes also referred to as a conditional wait). Delays are often dependent variables in studies using discrete-event simulation.
- **Clock:** A variable representing simulated time.

Robinson further classifies Events into two different types:(60)

- **B-Events:** Occurrences that are scheduled to occur at a specific point in time. These are usually arrivals of entities or completion of activities.
- **C-Events:** Occurrences that are dependent on conditions in the model as its state changes over time, usually related to the conditional start of some activity.

Discrete-event simulation evolves by the interaction of the above over time, in an ordered algorithmic way.

2.5.1.2. Basic Algorithm: Event Scheduling, Time-advance and System State

The FEL plays a central role in ensuring that events occur in the correct order. The basic approach is to keep taking event notices from the FEL, update the system clock to the time of the imminent event (the next event that will occur), execute the imminent event, change the system state and repeat the process. Because the FEL is maintained in a strict chronological ordering, the flow of events will always be in the correct order. The fundamental components of this algorithm are shown in Fig 2.1 (on the next page).

Each iteration of the algorithm shown in Fig 2.1 will result in a different system state, as recorded in state variables, counters, the system clock and the FEL. It is possible to execute this kind of algorithm by hand, writing down all of the changed state values in a table. Texts on discrete-event simulation

typically give such examples, showing in tabular form how system state and other variables change iteration-by-iteration in a simple model.(5,59,60)

The algorithm shown in Fig 2.1. will continue to execute until a stopping condition is met (the stopping condition is evaluated at step 8 of Fig 2.1). This is often specified as a future time, or alternatively as the occurrence of some specified event. For example, it may be decided in advance that a simulation should run for 30 days of simulated time in order to produce meaningful data. Alternatively, a specific event may be of interest in the modelling objectives (such as a queue reaching a specific length), and this is used as a stopping condition for the simulation. The problem of determining the length of a simulation run is important when considering the precision of estimates from the output data. This is considered in more detail in 2.5.3.4.

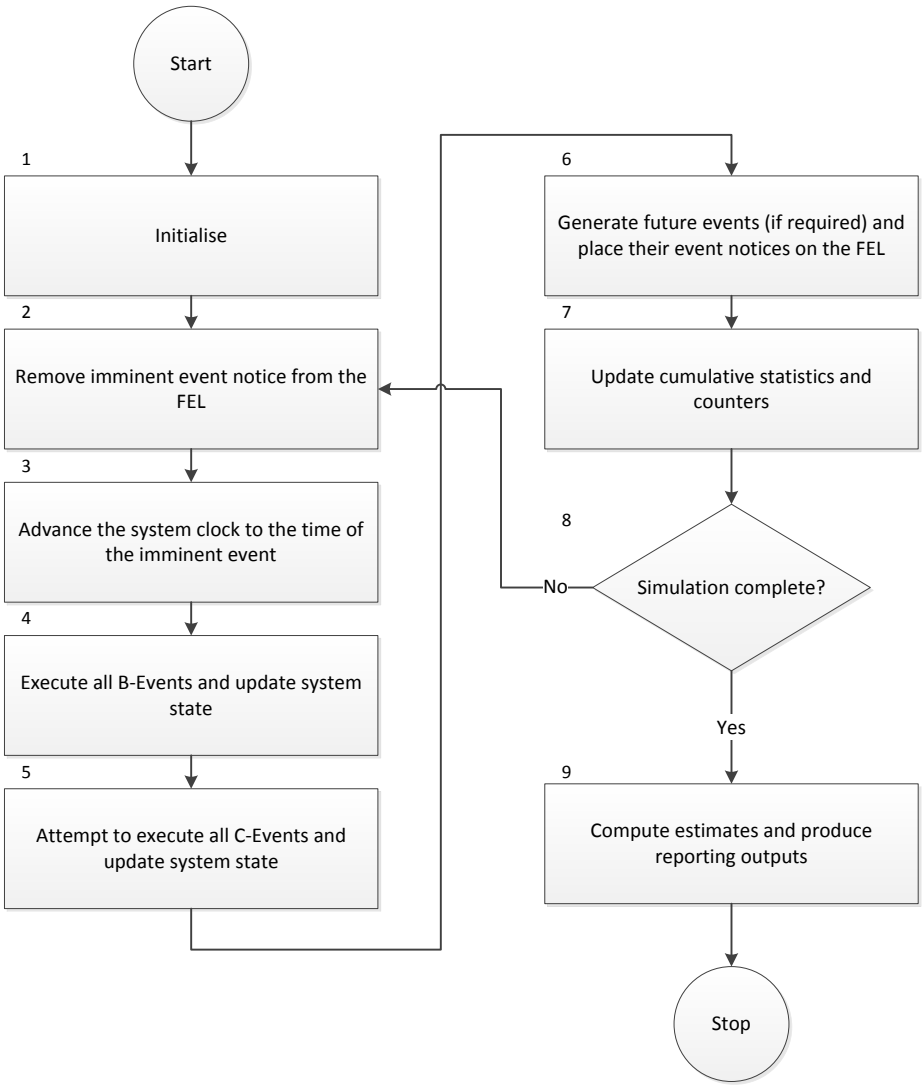


Figure 2.1. Basic Discrete-event Simulation Algorithm
FEL = Future Event List
Adapted from Robinson,(60) Law and Kelton(5) and Banks *et al.*(59)

2.5.2. Variability in Simulation Modelling

2.5.2.1. Random Numbers and Variables

Virtually all systems contain an element of variability in their processes. Examples include the arrival times of entities (customers, patients, vehicles or other objects) at specific locations, the time taken to complete various tasks or the failure rate of machines or other objects. Creating simulation models without the consideration of such variability would be unrealistic and would lead to output data which could not be used as an estimation of the modelled system's behaviour.(60)

Variability may be predictable (such as the on- and off-shift times of workers or machines) or unpredictable. Predictable variability is represented in a discrete-event simulation model by setting state variables to specific values at known times over the simulation time (for example by setting on- and off-shift times that repeat every day, or in some other predetermined pattern). Unpredictable variability is more complex to model, and requires the use of random numbers.(60)

Random variables are used in discrete-event simulation as a way of modelling system processes that are subject to unpredictable variability.(61) The principle is not simply to introduce any kind of random variation into these processes, but rather to model the nature of the system's variability as closely as possible. This requires access to data from the modelled system in order to determine the distribution and parameters of these data so that the model can be configured to generate random variables with as close a fit as possible. In the absence of reliable data from the system on which to base these decisions, distributions are often used which will generate approximately correctly distributed random variables based upon expert knowledge of the system.(62,63)

Once decisions have been made about random variables and how they are to be distributed in the simulation model, the model must be configured to produce these in an acceptable way. This is achieved using random number generators implemented in the simulation software that produce random number sequences imitating as closely as possible the desirable statistical qualities of random number sequences, namely uniformity and independence.(61) In reality, computer generated random numbers are not truly random – the term "pseudorandom" is used to describe them.(61,64) However algorithms exist that are capable of producing quality pseudorandom numbers that can be used to produce random variables following a wide range of useful statistical distributions.(61) Commercial simulation software products typically use one of these algorithms, a good example of which is the Mersenne Twister algorithm.(63,65)

2.5.2.2. Common Random Numbers

Most simulation software applications allow the user to specify a random number stream, in addition to a choice of probability distribution and parameters.(66) A stream can be considered an independent list of random numbers and the use of different random numbers streams in a simulation model is a popular method of variance reduction, the application of which is useful in increasing the precision of estimates when comparing different model configurations.(66–68)

If no random number stream is specified, a different stream will be used for each input distribution in the model, such as specific inter-arrival times or delays. If different model configurations are compared, this will lead to a situation where like input distributions across the configurations produce independent random variables. When random numbers streams are synchronised across different configurations, meaning that each input distribution uses an identical stream across configurations, the random numbers in these distributions may become positively correlated for each replication, which reduces the variance in estimation of output variables. This allows any real differences between model configurations to be detected with greater sensitivity.(67,68)

There is unfortunately no guarantee that the use of common random numbers will be successful as a variance reduction method. Adherence to guidelines for the implementation of common random numbers, for example as outlined by Law and Kelton and Banks *et al.*, will enhance the probability of a desirable effect.(67,68) The use of common random numbers as discussed above can be considered a part of experimentation validation together with the determination of a warm-up period, optimal run-length and optimal number of replications for experimentation (see 2.5.3.4).

2.5.3. Model Development in Discrete-event Simulation

A summary of the discrete-event model development process was given in Chapter 1, as an overview of the study methodology. The steps in this summary, with the exception of problem formulation, will be expanded upon below in order to give a more detailed description of key modelling concepts and processes, particularly where relevant to the current study.

2.5.3.1. Conceptual Modelling

Formulation of a simulation conceptual model is an essential step in discrete-event modelling.(1,69) The conceptual model is effectively a high-level, abstract, platform and software independent representation of a real or proposed system. Robinson defines a simulation conceptual model as a

non-software specific description of the simulation model that addresses six key areas: (i) objectives, (ii) inputs, (iii) outputs, (iv) content, (v) assumptions and (vi) simplifications.(69)

Simulation conceptual models may be non-formal (i.e. exist only in the modeller's mind) or may be formal and explicitly described. Regardless of which approach is taken, the conceptual model must serve a purpose and not merely be seen as a "checkbox" step to be complied with in the process of modelling. In an age of increasingly powerful simulation software, and software user interfaces that allow for rapid development of simulation models, the relevance of conceptual modelling has been questioned. However Robinson makes the point that more sophisticated simulation software, with the ability to create more complex models in shorter time frames, actually makes conceptual modelling more important than less so.(69) Spending time on this activity ensures that the resulting software representations of the model in question are focused on their original objectives, as simple as possible and feasible to implement within reasonable time frames.

The purpose and importance of a conceptual model goes beyond a focus on objectives and the avoidance of unnecessary complexity, however. It also plays a key role in clarifying, documenting and communicating fundamental structure and function of the model and articulates how the void between the simulation problem and the eventual software model implementation has been bridged. Robinson further adds that the conceptual model is important in forming the basis of subsequent verification and validation of the model and in making a case for credibility of the model (i.e. a perception that the conceptual model can be transformed into a software representation that is sufficiently accurate for the problem under consideration).(69,70)

Although the purpose for, and importance of, constructing a conceptual model prior to the initiation of software modelling is clear, there has been a long-standing lack of standardisation in approaches to conceptual modelling and frameworks for this activity in the literature. Many texts on the subject of simulation modelling mention the conceptual model as important, but provide little guidance on how to construct a conceptual model or what should be in it, saying only that development of conceptual models in more of an "art" than a "science" and is learnt through experience.(1,8,37) Robinson's text is has been chosen as a basis for conceptual modelling in the current study because it is a coherent and user-friendly source for information on conceptual modelling for discrete-event simulation and puts forward a useful framework to this end. Consequently, this framework has also been chosen as a way of structuring, describing and presenting the conceptual model in the current study.

Robinson's framework is described as a sequence of activities required to develop a simulation conceptual model. It has been created with the intention of filling a critical gap in this area (as described above) and is intended to be used primarily within the broad domain of operations systems discrete-event simulation. "*Operations Systems"* (also sometimes called "*Operating Systems"*) are defined as systems focused on combinations of manufacture, transport, supply and service and are "business-orientated", with Robinson pointing out that "business" is used broadly and includes both health care and the public service.(71)

As indicated above, in the first definition of a conceptual model, Robinson's framework is comprised of the following components, presented in more detail in points (i) - (vi) below.(71) A final element, "Process Logic" has been added. Although Robinson's framework does refer to the utility of diagrammatic representations of aspects of a conceptual model, this is not explicitly defined as a part of the framework. It has been added here as process logic of the model is crucial to its credibility and validity and is not suitably defined in any other part of the framework.

(i) **Objectives:** Determining the modelling objectives which establish the direction of all other modelling efforts and define the activities below (inputs, outputs and content).

(ii) **Outputs:** Identifying the model outputs, also referred to as responses (conceptually equivalent to dependent variables in experimental research design).

(iii) **Inputs:** Identifying the model inputs, also referred to as experimental factors (conceptually equivalent to independent variables in experimental research design). Inputs may also include probability distributions and other data used to drive simulations.

(iv) **Content:** Determining the model content, considered in terms of its scope (boundary or breadth of the model) and level of detail (depth or complexity).

(v) **Model Process Logic:** Description of the logical algorithmic execution of processes and decision points relevant to the model, by means of annotated flow diagrams.

(vi) **Assumptions and Simplifications:** Identification of assumptions (assumed truths stated in response to uncertainty or beliefs about the system being modelled) and simplifications (decisions made on the basis of their desirability in making the model more feasible to implement or understand, or to reduce input data requirements).

The conceptual model detailed in Chapter 3 of this study follows the framework structure presented above.

2.5.3.2. Data Collection

Data from the system being modelled are needed during two main stages of the model development process: (i) during model translation and (ii) during validation of the model (both are described in more detail in 3.6 and 3.8).(72) The kind and nature of data required vary depending on the type of model, its complexity and the modelling objectives. However in the first case (model translation), system data are typically needed in order to determine the distributions and other characteristics of key variables such as inter-arrival times of entities or processing times of activities. These quantities must be determined somehow in the software implementation of the model, and it is through the analysis of real system data that this is accomplished. In the second case (model verification), system data are required for comparative purposes; to compare the model's output with, as a way of determining whether the model is an accurate representation of the real system.(62,72)

Data needs are determined as part of the conceptual modelling process described above, although some preliminary data may be required before the conceptual modelling process begins, or early in it. During refinement of the model outputs and content (components (iii) and (iv) of the conceptual model framework presented above) input and output data needs should be determined. Data should be obtained from the system that have been recently generated and are not subject to the effects of processes that are different, or have undergone change, compared to the processes being modelled. In addition, the use of histograms and scatter plots to assess input data prior to further analysis may be helpful in detecting unusual distributions or unknown relationships between variables.(62)

One of the main aspects of input data analysis is the fitting of system data to a statistical distribution in order to sample random variables during the simulation run from the same (or a very similar) distribution. This is most effectively done using an input analysis software application. Applications of this nature typically will accept large volumes of input data read from a text or similar file and will provide descriptive statistical analysis of the sample including the generation of histograms. These applications allow the user to choose one or more goodness-of-fit tests (such as Chi-square or Kolmogorov-Smirnov) in order to test the hypothesis that the sample data follows a specific distributional form. Parameters for each fitted distribution are also computed. Input analysis applications typically allow users to assess the fit and determine parameters for a range of probability distributions commonly used in modelling.(62,63)

Although goodness-of-fit tests are useful in deciding on the best fit of sample input data to a particular distribution, they are only one factor to be considered in determining the correct choice. In particular, just like any other hypothesis test, goodness-of-fit tests are sensitive to sample size. Both small samples and very large samples can be problematic and may lead to incorrect conclusions being drawn based purely on the results of these tests.(62,63) Other factors to be considered are the shape of the histogram in relation to the shape of a particular probability distribution and the magnitude of the square error (reported by some input analysis applications) associated with each fitted distribution.(62,63,72)

Two cases in input data analysis require specific approaches. The first case is that of obtaining a large volume of data for input analysis (several thousand records for example). In such cases it is highly likely that goodness-of-fit tests will reject every candidate distribution. Consequently, an alternative approach must be followed in order to decide on the best fitting distribution. A smaller random sample can be drawn from the original data set and used for input analysis, or other approaches (as described above) can be taken to choose the best distribution.(63)

The second case is that of having no data for input analysis. This may arise because the system being modelled (or part of it) is only hypothetical and does not exist in reality, or because the specific measurement required is not measured or obtainable from the modelled system. A recommended approach in this situation is to utilise a distribution which has very few assumptions, or assumptions that can be based upon the opinions of system experts. If only an upper and lower limit for the input data range can be determined then a uniform distribution can be used.(62,63) If, in addition to this minimal information, the most likely value can be estimated without data then the triangular distribution can be used.(62,63) Alternatively, a beta distribution(62) or beta-PERT distribution(73,74) can be used. Of these, the uniform distribution is the least satisfactory and the beta or beta-PERT tends to be the most satisfactory.(62) Knowledge of the process being modelled can also be used to decide on a probability distribution when data are available, as some process types are known to be well represented by specific distribution properties.(62,72)

Once input data analysis has been completed, and decisions have been made about the best probability distribution for random variables, the model must be configured to produce the chosen values, proportions or distributions of random variables. This is done as part of the model translation process.

2.5.3.3. Model Translation

The process of translating a conceptual model into a software simulation model may take many different forms. In some cases, the model is coded in a general purpose or simulation programming language.(66) More commonly, a commercial simulation package is used for model building and simulation. Regardless of the approach, every component of the conceptual model must be translated into a software representation as accurately as possible and with adequate documentation of what has been done.(66) Details of the model translation process in the current study, along with those of the conceptual model, are set out in Chapter 3.

2.5.3.4. Model Verification and Validation

Verification

Verification refers to the process of ensuring that the conceptual model has been translated into a computer model (sometimes referred to as an operational model) with sufficient accuracy.(9,37,75,76) Accuracy in this context specifically refers to the degree of similarity between conceptual and computer models. A computer simulation that produces behaviour as specified by the conceptual model every time it is run is said to be correct.(37) Although verification is narrowly defined, it is more broadly speaking a component of model validation.(75) When considering the accuracy with which a computer model represents a conceptual model, it is important to note that this can never be complete. *Sufficient* accuracy must take into consideration the purpose of the model and must be assessed against this rather than applied in an arbitrary or absolute way. Thus, although emphasis is placed on verification as striving to show that a computer model is *correct*, this should rather be thought of as the accumulation of evidence that a model is *not incorrect,* which increases confidence in the model to the threshold of *sufficient* accuracy.(75)

Verification is considered to be an incremental, ongoing process during translation of a conceptual model, rather than a single procedure carried out once a computer model has been constructed.(9,75,76) In keeping with this, verification typically occurs as a computer model is being developed and involves smaller components of the model, through to more and more complex parts and eventually to the model as a whole. Although the definition of verification given above may suggest unidirectional movement from conceptual model to computer model as verification proceeds, this is not the case. New insights about the model uncovered during verification may require adjustment of the conceptual model and re-verification. Thus there is a continuous and progressive interplay between adjustments to the conceptual and computer models as verification proceeds in an incremental fashion.(75)

There is no strict checklist approach to model verification, however a number of methods have been suggested that all may be used to some extent in the verification process depending on the model characteristics and software or programming environment used.

(i) Modular Development and Testing

Law and Kelton recommend developing a simulation model in small, unitary modules and testing each one thoroughly.(8) This way, a situation is avoided where a large amount of complex logic is created and only tested at the end of the development process, making the source of errors difficult to identify. Once the fundamental modules of a model have been developed and tested, they can be logically "joined" into larger functional units until, following a layering approach, the final model is arrived at. Although the description of this process is really aimed primarily at model development by programming, the principles and benefits apply equally to models developed using commercial simulation software. Modular development and white-box validation (discussed below) refer essentially to two aspects of the same process.

(ii) Visual Checks

Visual checks are carried out during modelling and may apply to checking of static parts of the model, such as configuration details and settings, data tables, process logic and state variable configuration on initialisation.(8,9,75) Visual checks can also mean that the model is observed dynamically as it runs during development. This is typically done in a start-pause fashion, allowing for checking of various parts of the model and the behaviour of objects as execution of the model is advanced in small increments.(9,75) There is some overlap with (iv) and (v) below, as this process is greatly facilitated by having a quality animation of the model to observe and by having a detailed trace of the model's execution to check.

(iii) Checking Output

The model's output data, in the form of reports at the end of a simulation run or specific state variable values during a run, can be assessed and checked to see if their values are in keeping with what is expected, given the nature and stage of the model and its input data.(8,9,75) Reports that include confidence intervals or other measures of dispersion can be useful in detecting processes which may not be producing data in the way they are expected to. Likewise, using charts (mainly line charts, scatter plots or histograms) of output data can very effectively and quickly show extreme or outlying data, or trends that indicate problems in process logic or resource utilisation.(75,76)

(iv) Use of a Trace

A trace is a step-by-step record of the simulation's execution, containing information about the simulation time, state and all entities and processes in the simulation at a given point.(8,75) A trace can be viewed when the model is paused during execution of a run. Some simulation products save the trace in a text file on disk, which can then be opened and viewed after the simulation run has been completed. The trace is actually a form of output report (as in (iii) above) and is used in much the same way for verification purposes. The main advantage of using a trace is the amount of detail that can be obtained from it, and the ability to reconstruct complex processes and interactions in the model step-by-step. Invariably, if abnormal behaviour is observed in a simulation run the trace must be consulted in order to determine exactly why the behaviour occurred.

(v) Observation of Animated Behaviour

Many commercial simulation products are able to provide an animated interface through which behaviour of the model can be observed as it runs (this relates to (ii) above). Observing the model's behaviour and comparing it to known or required behaviour of the system being modelled is a very useful way of identifying logic errors that may not be easily identified in other ways. In some cases a "Watch" facility is also available. This means that a simulation run can be paused at any point and any object or state variable can be observed with the values and configuration it has at that moment. Use of a Watch facility is useful when abnormal behaviour has been observed during animation, and fine detail about the entities and state variables involved is needed for diagnostic purposes.

(vi) Use of a Commercial Simulation Product

Law and Kelton suggest that the use of a commercial simulation product, rather than coding of a simulation model "from the ground up", should be considered a factor enhancing the ease of model verification as the complexities and pitfalls of coding are avoided. They warn, however, that care should be taken in the choice of a commercial product and attention given to its credibility and reputation, as well as how adequately it is documented.(8)

Validation

Validation is the process of assessing whether a simulation model is accurate enough for the specific objectives of a study.(8,75) Banks *et al.* differentiate validation from verification by describing it as "building the correct model", whereas verification is concerned with "building the model

correctly".(9) Although verification and validation are often defined and described separately Robinson points out that, broadly speaking, activities concerned with verification can often be seen as a subset of those concerned with validation.(75)

Both verification and validation are performed in an iterative way, throughout the life-cycle of a simulation study (8,9,75) and the term validation can be used in a more specific way to refer to validation of the conceptual model, validation of input data, and so forth. Validation of the conceptual model is concerned with assessment of whether the content, assumptions and simplifications of the model are sufficiently accurate, given the objectives of the simulation study. This is normally determined by consultation with system experts and users, or by comparing the conceptual model with system documentation.(75) Validation of input data focuses on assessment of the accuracy of system data for its use as model input.(75) A variety of methods mentioned in 2.5.3.2 are used for input data validation. Three further specific types of validation are discussed in more detail below.

(i) White-box Validation

White-box validation is a small-scale or micro check of a part of the model's validity.(75) This is typically done as constituent components of the model have been developed. Verification is performed in parallel with white-box validation, however verification is concerned with the relationship between the model component and the conceptual model, while white-box validation is concerned with the relationship between the model component and the real system.(9,75) Similar methods may be used though, and these are largely centred on observation of behaviour and analysis of output data.

(ii) Black-box Validation

Black-box validation is a macro or complete model-scale check of validity. In other words, black-box validation (referred to as *results* validation by Law and Kelton) is concerned with an assessment of the behaviour of the whole model and whether or not this is accurate enough for the objectives of a specific study.(8,75) This is typically done by comparing outputs of the model, primarily those of interest in the application of system design (or the dependent variables of the study), with those of the modelled system.(8,9,75) In more limited circumstances, it may also be possible or even required to compare the model to be validated with another model such as a mathematical model.(75)

The mainstay of black-box validation is quantitative comparison of model and system outputs. This is not as simple as performing a hypothesis test to compare means, as data from virtually all real systems and simulations are non-stationary and autocorrelated.(8) Consequently, the normal statistical tests based upon independent and identically distributed observations are not valid. Law and Kelton further point out that that the null hypothesis (literally "no difference") as traditionally stated in hypothesis tests comparing means is also not valid in the context of a system-model comparison because a model is not expected to be anything more than an approximation of a real system.(8) For these reasons, the following approaches are recommended in quantitative comparisons of system and model data:

Inspection: Comparison of system and model output data without the use formal statistical procedures, using point estimates (such as the mean and variance) and charts. Although widely used, this approach is subject to randomness of the observations and can lead to the incorrect conclusions being reached. As an alternative, the model can be run with historical input data (referred to as a trace-driven simulation) and compared with the same set of observations as the system output.(8)

Confidence Intervals: Confidence intervals can be used to compare the mean difference between system and simulation output data, providing as much information about equality of means as a traditional hypothesis test, but also giving an indication of the magnitude of the difference. Two modified methods for constructing confidence intervals for model validation can be used, each with different assumptions.(8)

As an additional non-quantitative validation method, a Turing test may be performed.(8,75) This involves providing system experts with output from both the system and the model, presented separately and in the same format. If it is possible for system experts to differentiate between the two sets of output, the basis of their discrimination can be used to improve the model. If, on the other hand, they are not able to differentiate between system and model output data, this can be used as evidence of the model's validity and credibility.(8,75)

(iii) Experimentation Validation

Experimentation validation is concerned with the question of whether the experimental procedures used are accurate enough for the study's objectives. This is considered in three main sub-questions relating to whether a warm-up period is required in each simulation run (and if so, what it should

be), how long each simulation run should be and lastly, how many replications should be performed in an experiment.(75)

The starting point for addressing the above questions is whether the simulation is terminating or non-terminating. A terminating simulation is one where a run of the simulation reaches a natural end point. This can be determined a number of ways, but the most obvious case is a simulation of a system that cyclically reaches an empty state. A good example of this is a system where customers are serviced strictly during business hours, such as a bank. In contrast, a non-terminating simulation is one that has no natural end point, obvious examples of which include production facilities or other services that run continuously.(77,78) Emergency services, whether medical or of any other nature, are typically examples of non-terminating systems.

Non-terminating simulations will typically produce initial transient output data from start-up, meaning that the distribution of the output is constantly changing. After some variable time period, such a simulation will generally produce steady state output, characterised by variation in accordance with some fixed distribution. Steady state does not mean that values of a particular random variable are all the same for a given simulation run, but rather that the variation in such a random variable follows a fixed distribution.(77,78) In some cases, such as those where a time period is divided into shifts where processing or arrival rates differ, a steady state cycle may be apparent (effectively two steady states, with cycling between them).(77) In such cases, if observations are taken in a period as long as the longest cycle then analysis can be performed as if the output is steady state.(78) The importance of identifying if and when output data achieve a steady state relates to the fact that data recorded before this point should not be included in analysis due to initialisation bias.(77,78)

Two methods may be used to avoid initialisation bias. The first is to set initial conditions in the model at the start of a simulation run thus immediately placing the model in a realistic condition. The second method is to run the simulation for a warm-up period, the duration of which is considered long enough for the simulation output to be in a steady state. The determination of this time period (i.e. the warm-up time) may be approached a number of different ways including the use of time series and other charted output, heuristics and statistical methods. One graphical method devised by Welch and based upon the calculation and plotting of moving averages to identify the onset of steady state, is widely recommended according to Robinson.(78)

After consideration of measures to avoid initialisation bias two further questions remain, namely how long each simulation run should be and how many replications of these runs will be required in order to produce data of sufficient quality for experimentation. “Sufficient quality” in this context refers to data capable of accurately estimating the model’s performance in terms of some set of chosen output variables.

Run length may be determined by characteristics of the modelled system and by the objectives of the simulation study. In terminating simulations, run length may be equivalent to the duration of system activity or processing. In non-terminating simulations, run length is not subject to considerations of natural end points, and is typically guided by study objectives and requirements related to the range of conditions under which the model is to be observed (which vary over time, often is some cyclical way). Additionally, in the case of non-terminating simulations where multiple replications are utilised, a rough rule is to set the replication length to 10 times the warm-up period although other considerations may shorten or lengthen this value.(79)

Selection of an appropriate number of replications is determined in a more quantitative manner. A replication is a single run of the simulation using a specific stream of random numbers. The first consideration is the inappropriateness (in most cases) of drawing conclusions about system performance from a single replication, as highlighted by Law and Kelton.(77) Because of the variance in random variables selected from a given distribution between individual runs of a simulation, estimates based on a single simulation run could be quite different from those of other runs and from those characterising the true performance of the model. Each replication is equivalent to an independent random sample of output data from a system, and thus estimates must be taken from a set of replications rather than just a single one. In the case of terminating simulations, there is no alternative but to use multiple replications. However in the case of non-terminating simulations, it may be more intuitive to use a single long run. In either case, confidence intervals must be calculated as a way of expressing estimates of output variables which poses a particular challenge in the case of a single long run. For this reason, it may be appropriate or even preferable to use multiple replications in the case of non-terminating simulations, together with longer runs.(78)

Two main methods are described that can be used to guide the decision on how many replications to use. The first involves plotting a cumulative mean of output variables of interest, based upon an initial selection of 10 replications. An estimate of the required number of replications is made based on the point at which the cumulative mean line becomes flattened. An alternative is to use

confidence intervals calculated for each replication. The principle is similar to that in the previous approach, except that more information is available on the precision of output data as an estimate from the confidence intervals. For each replication, the deviation from the mean of each confidence interval can be calculated and a maximum value for this can be used as a decision point for the required number of replications, based on the objectives of the study. In both of the above cases, output data from the warm-up period are deleted before analysis.

2.5.3.5. Experimentation and Data Analysis

Experimentation, or more specifically *batch* experimentation, is carried out by applying the settings for warm-up (if required), run length and replications discussed above together with experimental factors. The simulation is then run to produce output data which is then analysed in accordance with the study objectives which may be to analyse data from a single scenario or to compare alternative scenarios or even models.(80)

A number of approaches to statistical analysis are described that apply specifically to simulation output data. In some cases, traditional statistical methods for independent samples can be used, however in other cases the nature of the data from a simulation necessitates adaption in the method used for calculation of confidence intervals, in the case of either single or multiple comparisons.(80,81) More complex experimental designs can be utilised in assessing the main effects and interactions of experimental factors and a range of optimisation techniques can also be applied in searching for a "best" combination of experimental factors to maximise or minimise a response.(80)

2.6. SIMULATION-BASED STUDIES OF EMERGENCY MEDICAL SERVICES RESPONSE SYSTEMS

EMS in the broader sense, and specifically the performance of EMS response systems, has been the focus of research efforts for several decades. Approaches to the simulation modelling of EMS systems have been reported in the literature from the late 1960s, with an increase of this material in the last 10 years.(82) Increasing interest in this area of research and its applications may be related to the development of EMS as a whole and the quest for more sophisticated methodologies to apply to systems design and operation, as well as increasing emphasis on performance optimisation in the face of growing competition for health care funding.

A few examples exist of performance measures related to patient survival in the simulation and modelling literature,(83–85) however virtually all published studies to date have been devoted to

some aspect of timeliness as a measure of system performance, with the related problem of vehicle utilisation being almost as prevalent.(82) Of these, 25 published works have been identified that deal primarily with mean response time, or coverage of a geographic area within a specific response time, as a performance indicator.

2.6.1. Modelling Approaches Used in Studies Involving Response Time

A wide variety of modelling approaches are used in the studies identified above. Modelling applied to different features of the simulation in each of these studies is discussed under the sub-headings below.

2.6.1.1. Demand and Dispatch Modelling

Demand Modelling

The modelling of demand in EMS systems is approached in two main ways; modelling of rates and distributions of arriving calls[†] and modelling of geographic or spatial distribution of calls.(82) Most studies used a Poisson distributed call arrival rate,(86–88) with some variations such as adjustment for demand variations as a function of time of day(89–96) or geographic location.(97) Other approaches included empirical distributions(98) and in two cases, historical call arrival data were used to drive the simulation.(99,100) In one case a spreadsheet-based modelling approach was used, using a Poisson distribution for every two hour interval over a week of sampled data.(101)

Spatial modelling of call distribution utilised two main approaches; pre-existing areas and areas determined specifically for the purpose of the simulation. In the former case districts, regions, census tracts or postal (zip) code areas were used.(88,90,93,96,98,102) In the latter case, custom-defined areas referred to as zones, grids or nodes were defined based either upon dimensions that equally divided an area of interest or based upon areas of equal population.(89,91,92,94,97,99,100) Determination of call arrival rates within each of the above areas was accomplished by either assigning each area its own arrival rate and distribution (including Poisson, uniform or other random distribution) or by using empirical distributions or real call locations (mapped to nodes in the model), as indicated above.

[†] The term “call” is used in this part of the review as it is the term most often used in the literature. In other parts of this document the term “incident” is used. The two terms mean the same thing and can be used interchangeably.

Dispatch Modelling

Two major considerations are relevant to the modelling of vehicle dispatch processes. The first is the choice of vehicle to be dispatched to a call based purely upon attempting to optimise the response time, which takes into consideration factors such as distance from the vehicle's dispatch point to the call location or estimated travel time to the call location. The second consideration is matching the patient's condition or acuity with available vehicles and making a decision as to the best choice of vehicle to allocate, incorporating distance and travel time.

A variety of approaches were used for vehicle allocation (without consideration of call priority), however the most common method involved allocation of the closest vehicle to the call.(88,90,95–100,102–105) In only two studies was this basic approach modified with the addition of pre-emption, meaning that a vehicle *en route* to a low priority call could be diverted to a higher priority call if it was closer.(92,106) In some studies dispatch was determined primarily by regional allocation (91,93,107,108) and in one case the estimated travel time of each candidate vehicle to a call location was used for allocation decisions.(87)

Some studies took a simplified approach to dispatch modelling and did not differentiate between levels of patient acuity (or priority).(86,87,89,90,95,97,98,102) In such cases, either all calls were considered to be life-threatening or no specific reference was made to the acuity level of the call. Of those studies where calls were prioritised, a range of different schemes and terms were used for this. The basic approach however was to place calls into either two (life-threatening or non-life-threatening) categories, or more categories reflecting a broader range of states between these two extremes.(92–94,96,99,100) Where prioritisation was part of the dispatch process, vehicle allocation took this into consideration and in one case it was also factored into modelling of scene times and hospital handover times.(92,94)

2.6.1.2. Travel Time Modelling

Considering that the studies reviewed here all identified response time as at least one of the outcome variables, the modelling of vehicle travel time is of obvious importance. Approaches varied in complexity, depending on the availability of data that could be used in this process. Examples of studies where a large amount of detailed data were available include those by Henderson and Mason,(100) Goldberg *et al.*(98) and Aboueljinane, Jemai and Sahin.(96)

Henderson and Mason had access to a travel time model capable of computing travel times between any two locations at any time, factoring in the effects of traffic congestion. Although this model was not used to compute shortest paths during the simulation, it was used to pre-compute these paths from all possible sources to all possible destinations for various travel times. Travel times during the simulation were then derived from travel along these paths.(100) Goldberg *et al.* factored different types of roads, and the relative distances travelled on each along the path to a call, in determining travel times.(98) Aboueljinane, Jemai and Sahin obtained pre-computed average travel times of all routes between zones used in their model. When combined with Global Positioning System data from real vehicles, routes from actual calls were identified and average travel times for every section of the road network were determined using a similar road type classification as used by Henderson and Mason. Consequently, travel times could be determined for every day, hour, call type and priority by combining these data with average speed data from vehicle Global Positioning System records.(96)

In other studies, travel time modelling was comparatively much simpler and made use of deterministic speeds and a variety of methods to calculate distances between vehicle and call locations, the most common of which was the Euclidean distance.(87,89) Correction factors were sometimes used to account for the effects of traffic or varying road types, both of which may affect travel time.(87) Speed was also computed as a function of distance, with calculations varying with the distance involved (shorter or longer distances).(94) Deterministic speeds together with the use of shortest path algorithms applied to a road network were also used.(92)

2.6.1.3. Process Times and Destination Selection Modelling

Processes included in EMS simulation models take account of typical activities occurring at various stages of the emergency response, including call taking and the selection and dispatch of one or more vehicles, on-scene activities that occur once the vehicle has arrived, transportation of one or more patients to hospital and hand-over at the receiving hospital and preparation for the next call.

The approach used in modelling the time intervals for these processes was largely dependent on whether or not adequate system data were available for input analysis. In several cases adequate data were not available, and deterministic approximations were used for on-scene, handover and preparation times.(87,94,95) In two cases, the time taken to dispatch a vehicle once the dispatch centre had been notified was considered to be negligible and was assigned zero time in the model.(89,107) In other cases, data on process times were available for input analysis for some or all

of the processes of interest and the usual approach of fitting probability distributions to these data for use in the model was followed.(92,93,107)

With regard to the selection of a destination hospital once on-scene processes were completed, two main approaches were taken in modelling. The first approach was simply to select the closest hospital to the call location.(88,89,91,99,107) The second approach was to factor other considerations into the selection process such as whether a given candidate hospital had capacity to receive patients or whether the appropriate resources for a given patient were available at a specific hospital.(94,100,106) In such cases empirical distributions of hospital selection were used, or the simulations were trace-driven.(100)

2.6.2. Verification and Validation of Studies Involving Response Time

Several approaches described in the simulation and model-development literature and outlined earlier in this Chapter were used to validate simulation models in the chosen studies. Some of these are described above as verification methods, however as Robinson points out,(75) most of these are common to both verification and validation. Examples of some of the approaches used include modular implementation,(87,103) the extensive use of traces during development(94) and visual checking of model behaviour by observing animations.(100) The opinions of system experts were also sought in some cases in order to assess model accuracy.(87,107) Black box validation was used, incorporating a mix of both inspection of model and system output data, and more formal statistical comparisons of either output variables or distributions.(92–94,96)

2.6.3. Explanatory Factors of Studies Involving Response Time

The studies discussed above all included response time as a dependent variable, either in the form of mean response time or expressed by consideration of a spatial coverage area serviced within a specified response time. However there was a great deal of variation within this grouping of independent variables or explanatory factors. The following sub-headings are used to arrange these into a number of categories.

2.6.3.1. Location and Placement of Vehicles

Location and placement of vehicles refers respectively to where vehicle bases should be located and to where vehicles should be placed if not at bases, for example at holding or waiting points. In some cases these two questions refer to the same thing (i.e. vehicles only wait at bases for calls). The earliest example of such a study is that by Savas which investigated a range of factors on response

time, including the effect of a new base (located in an area of high-demand) and more vehicles.(106) Others have focused on similar problems, devoted to assessing the effect of either new bases or the allocation of more personnel and vehicles to existing bases.(94,96,106) Decisions used for the placement and location of vehicles ranged from those based on heuristic approaches(89) to those based on probabilistic(86) or deterministic mathematical models.(97,99,103)

2.6.3.2. Redeployment of Vehicles

The simplest case of vehicle location is the case where vehicles wait for calls at a static position, meaning that the position does not change in the short term. More sophisticated approaches to vehicle location, or deployment as frequently referred to in the literature, take into account the very short-term shifting nature of demand in a geographic area of service driven by activities, traffic patterns, seasonal influences and other factors over periods of hours, days or weeks. Redeployment is considered as being either multi-period, which is deployment to address changing demand patterns, or dynamic, to solve the problem of where to base vehicles when they become idle over a shorter time-frame.

Decisions regarding exactly how to choose a redeployment policy for vehicles can, like the placement and location decisions, be made heuristically or analytically with mathematical models. In one example, Geographic Information System data were used to define coverage areas for a given maximum response time and vehicles were placed in these areas to solve a multi-period redeployment problem.(109) Equity in the distribution of vehicles over an area was considered in another approach which was used to compare solutions to a dynamic redeployment problem (94) and several dynamic redeployment policies have been compared in a study from the Netherlands.(88) Mathematical models for vehicle redeployment decisions in simulation have been used less frequently.(93,95)

2.6.4. Factors Found to Decrease Response Time

In the studies discussed above, a number of factors have been identified as having a favourable effect of response time. One of the earliest studies of simulation application to an EMS problem suggested that location of vehicles at a position within an area of demand (as opposed to a hospital) improved response time. This study also found that adding vehicles to the original location (the hospital) had an initial, relatively small beneficial effect on response time which soon reached a plateau and that this effect was due to the reduction in waiting time (the time taken to allocate a call to a vehicle) with increased vehicle numbers.(106)

The location and placement of vehicles, which was a focus of investigation in many of the studies, was found to have a beneficial effect on response time.(86,89–91,93,94,96,97,99,102,103,110) However it is difficult to extract general patterns from these studies as the specific questions asked in each case, and the solutions investigated, were unique to each modelled system. Some of the redeployment policies investigated in three studies were shown to decrease response times.(88,95,109) A few other studies investigated changes to vehicle numbers in a given area, finding that in some cases improvements in response time could be demonstrated although these were mostly quite limited.(87,89,94,97,99,102) A single study showed that decreasing the time delay for dispatching vehicles improved response times,(96) while another demonstrated the effect of increasing vehicle average speed on response time in high and lower acuity cases.(99)

2.6.5. The Use of Discrete-Event Simulation

Many different simulation approaches were taken for all of the studies discussed above. In some cases the method used was some form of simulation alone, while in other cases analytical methods were used to solve a specific part of the problem and this was used together with simulation to model a given system and generate data for analysis. Of the 25 studies selected, which included response time as a dependent variable, 11 (44%) utilised discrete-event simulation.(87–89,94–97,102,107,108) Where discrete-event simulation was not specified explicitly as the simulation method used, the approach tended to be custom-developed simulation software which most likely still used a discrete-event methodology.(98,100) Several other examples not included in the above discussion were found of discrete-event simulation applied to EMS problems.(111–114)

2.7. SUMMARY

The literature reviewed in this Chapter was focused in two main areas. The first was on the role of response time as an EMS performance indicator, taking into consideration the relationship between timeliness as a criterion of EMS quality and the performance indicator of response time as a measure of this criterion. The meaningfulness of response time as a performance indicator was considered critically, leading to a conclusion that although there is not strong evidence to support the effect of response time on clinical outcomes, it is still regarded as at least one important EMS performance indicator at the present time. This underscores the rationale for studying the effects of EMS system design factors on response performance.

The second area of focus in this literature review was on computer simulation, and more specifically discrete-event simulation, arising from the choice of simulation as the methodological approach for this study. After a description of the basic conceptual underpinnings of discrete-event simulation, a description was given of the key processes and practices relevant to model development. The final sections of this review connected the principles of simulation back to the area of EMS performance by discussing studies from the literature where simulation was used as a method for investigating EMS system operations. Specific attention was given to studies concentrating on response time as an outcome variable. Information presented in this latter part of the Chapter sets the scene for a detailed description of the modelling approach in this study, presented in Chapter 3.

CHAPTER 3: RESEARCH DESIGN AND METHODS

3.1. INTRODUCTION

In preceding Chapters, the background and context of this study have been established along with its aim and objectives which focus on the effects of three experimental factors - ESV location, response model and ESV numbers - on response performance in a large, urban EMS system in South Africa. As argued in Chapter 1, and as supported by the literature presented in Chapter 2, computer simulation is a viable method for the investigation of EMS system response performance and was thus the method of choice for this study.

This Chapter is devoted to a description of the research design and a detailed account of the research method. The conceptual model is first presented, using as its structure the framework proposed in Chapter 2. This is followed by a description of the model's input data including sources of the data, descriptive analysis and probability distribution analysis.

Translation of the conceptual model into a computer model is presented by first describing the simulation software used, followed by explanations of each of the model's objects as represented in the computer model. This is followed by a description of the verification and validation procedures used in order to ensure that the model translation process was conducted accurately, and that the resultant computer model was a sufficiently accurate representation of the real system given the aim and objectives of this study.

The validated model was changed in order to implement system characteristics for each level of two experimental factors – ESV location and response model. These changes are described in the last part of the Chapter, followed by an account of the data analysis methods used in order to investigate the effects of ESV location, response model and ESV numbers on response performance.

3.2. RESEARCH DESIGN

An experimental research design was chosen for this study. This design is appropriate for the aim and objectives of the study, which are to establish a cause-and-effect relationship between changes in the selected independent variables and the dependent variables of response time and the proportion of responses meeting total response time targets.(115,116) Experimental research designs are characterised by two main requirements:(116)

- Manipulation of one or more independent variables, the response of which is measured and analysed with inferential statistical procedures.
- A high level of control over experimental and data collection processes, in order to limit the effects of extraneous variables.

These requirements have both been met in the current study as the independent variables of ESV numbers, dispatch model and geographic location of ESVs were directly manipulated in different formulations of the simulation model in order to assess their effect on the dependent variables. A high level of control was possible as each experiment was constructed using the same incident rate tables and each simulation run was executed for the same time period, yielding similar simulated system conditions.

3.3. SIMULATION PROBLEM STATEMENT

Unlike simulation aimed at solving a particular problem in a commercial environment, this study's objectives are directed towards answering questions about response system design choices in a more hypothetical way. The simulation problem is thus closely related to, and encapsulated in, the research problem as stated in Chapter 1. In essence, urban EMS systems in South Africa generally do not meet the response time performance goals specified at national level.(31–33) This may be due to a wide range of explanatory factors, some of which are likely system design factors. Three possible system-related explanatory factors have been selected for investigation, as described in the modelling objectives set out in 3.4.2.1.

3.4. CONCEPTUAL MODEL

3.4.1. Choice of Conceptual Modelling Framework

The conceptual modelling framework proposed by Robinson has been discussed in 2.5.3.1 of Chapter 2. In that Section, reasons were given for the choice of using Robinson's conceptual modelling framework in the current study. For ease of reference, the components of this framework are listed again below, before each of them is dealt with in greater detail. The components are:

(i) **Objectives:** Determining the modelling objectives which establish the direction of all other modelling efforts and define the activities below (inputs, outputs and content).

(ii) **Outputs:** Identifying the model outputs, also referred to as responses (conceptually equivalent to dependent variables in experimental research design).

(iii) **Inputs:** Identifying the model inputs, also referred to as experimental factors (conceptually equivalent to independent variables in experimental research design). Inputs may also include probability distributions and other data used to drive simulations.

(iv) **Content:** Determining the model content, considered in terms of its scope (boundary or breadth of the model) and level of detail (depth or complexity).

(v) **Model Process Logic:** Description of the logical algorithmic execution of processes and decision points relevant to the model, by means of annotated flow diagrams.

(vi) **Assumptions and Simplifications:** Identification of assumptions (assumed truths stated in response to uncertainty or beliefs about the system being modelled) and simplifications (decisions made on the basis of their desirability in making the model more feasible to implement or understand, or to reduce input data requirements).

3.4.2. Conceptual Model

3.4.2.1. Modelling Objectives

The modelling objectives were:

i) To create a baseline model of an urban EMS system based on empirical data and system characteristics from the Western Cape EMS operations in Cape Town.

ii) To alter the baseline model referred to above in order to create four different models representing all possible combinations of two factors, namely (ii) response model and (ii) ESV location, as described in 1.6.2 of Chapter 1.

iii) To determine which of the models in i) and ii) above produced the shortest response times and greatest proportion of P1 response times within 15 minutes and P2 response times within 60 minutes with a baseline number of ESVs, and to determine whether the identified model meets current response time benchmarks, as described in 1.6.2 (iii) of Chapter 1.

iv) If the identified model in iii) above did not meet the response time benchmarks, to determine the minimum ESV numbers required in the model to meet these benchmarks.

The rest of this conceptual model description will focus on the baseline model of the Western Cape EMS, Cape Town response system referred to in i) above. Differences in model process logic between this model and the four models representing combinations of the explanatory factors will be given after assumptions and simplifications of this baseline model have been presented.

3.4.2.2. Model Inputs and Outputs

Inputs (Experimental Factors)

- Response model (two levels: single- or two-tier)
- ESV location (two levels: static or dynamic)
- ESV numbers

Outputs (Responses)

- Response times (P1 and P2)
- Proportions of P1 and P2 cases meeting response time targets (see 3.4.2.1 above)

3.4.2.3. Model Content

Model Boundary

A high-level model of the interaction between the caller, patient, EMS and further care (hospital) is shown in Fig 3.1 (on next page). Following from the modelling objectives and the model's inputs and outputs as defined in 3.4.2.1 and 3.4.2.2, the model boundary is defined and depicted as the shaded area in Fig 3.1. The focus on response time as an output variable makes the caller (and time of call) a logical boundary of the system as an entry point, as response time is measured from the time the caller makes contact with the dispatch centre.

Although in a more holistic sense, and from the patient's perspective, an interaction with the system being modelled would conclude at the end of definitive care, this last phase has not been included in the model. Considering the modelling objectives, the exit point of the model boundary is where the last interaction influencing ESV availability and response performance occurs, namely after handover and return to availability of ESVs at the EC. This is depicted by the shaded area including the EC, but relates only to the process of handover to EC staff and making the ESV available and not any further processes in the EC. The blocks in Fig 3.1 represent objects of interest in the model, as bounded. These objects are represented as either entities, resources or other types of objects in the detailed model of the system.

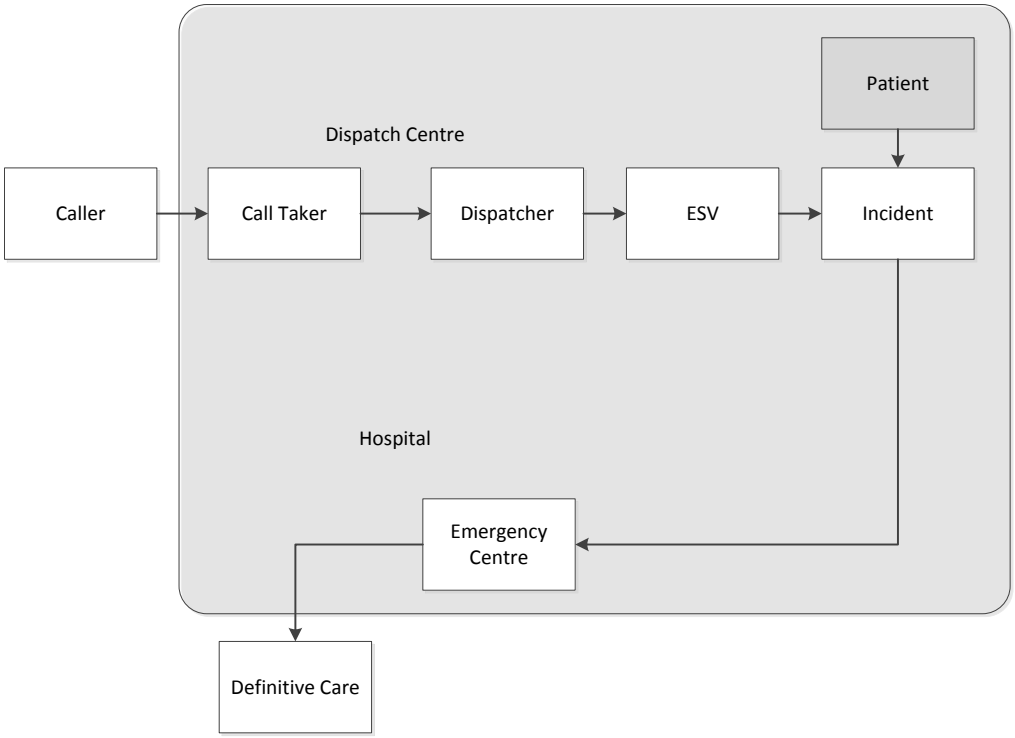


Figure 3.1. Model Boundary
ESV = Emergency Service Vehicle

Model Scope and Level of Detail

The model scope is defined by those components falling within the model boundary. The aim of simulation is to model as much of the real system as is necessary to address the simulation problem and not to attempt to model every possible aspect of the system. Accordingly, the level of detail relates to how much detail is needed in each of the included components in order to meet the modelling objectives. In addition to scope and level of detail, which are conveyed in Table 3.1 (beginning on the next page), a short description of each component is given.

Table 3.1. Model Components and Related Detail

Component	Description and Detail
Entities	
Patient Entity	*Description:* A patient entity represents a patient experiencing some kind of condition requiring EMC. For the purposes of this simulation, the exact nature of the condition is not important although patient entities are classified into two categories of acuity determined by their condition; Priority 1 (P1) and Priority 2 (P2). Patients falling into the P1 category require a more immediate, rapid response and those falling into the P2 category are considered to be less urgent. Typically, P2 patients are queued behind P1 patients when considering the allocation of finite response resources. Response system performance goals emphasise timely response to P1 patients, with response to P2 patients fitting within less rigorously defined limits. The prioritisation of patient entities is important in the current model because it is the primary consideration in determining the type of response and allocation of resources. Although all incidents and their related patient entities warrant a response, not all patients are transported from the incident location to a hospital. *Detail:* ▪ Arrival Pattern ▪ Attribute: Sector ▪ Attribute: Priority ▪ Attribute: Transport Decision ▪ Routing (if transported)

(Continues on next page)

Table 3.1. (Continued)

Component	Description and Detail
Resources	
ESV	*Description:* An ESV resource represents an abstraction of a vehicle that could be used for emergency response. Two main types of ESVs are available: (i) ambulances and (ii) PRVs. Ambulances are a resource which can be used for both emergency response and transportation of patient entities to their destination (one of a set of hospitals). PRVs are a resource which can be used only for emergency response. PRVs have no capacity to transport any patient entity and are thus reliant on ambulances to facilitate transportation of patient entities having been allocated their capacity. Ambulances may also be classified according to the level of care associated with the ambulance crew staffing them. Only a broad distinction is made here between ALS and non-ALS level of care, as this is a consideration during dispatch of ambulances to different priorities of patient entities (ALS ambulances being preferentially dispatched to P1 patients and non-ALS ambulances to P2 patients). *Detail:* ▪ Attribute: Sector ▪ Quantity (per sector) ▪ Attribute: Type (ambulance or PRV) ▪ Attribute: Level of Care (ALS or Non-ALS) ▪ Attribute: Transport Capacity ▪ Routing (derived from patient entity routing) ▪ Attribute: On/Off Shift

(Continues on next page)

Table 3.1 (Continued)

Component	Description and Detail
Other Objects/Components	
Dispatch Centre	*Description:* A dispatch centre is the first point of contact for a caller reporting an emergency of some kind. The caller may or may not be the patient. The dispatch centre is broadly divided into a call taking section and a dispatching section. For the purposes of this model, very little detail is needed in any representation of the dispatch centre and it is not necessary to separate call taking and dispatch. The importance of the dispatch centre in relation to the modelling objectives is that it represents a variable time delay in processing of a response, and thus will affect response time. *Detail:* ▪ Attribute: Processing Delay
Incident Location	*Description:* An incident location is a position in two-dimensional space where the emergency (also referred to as an incident) has occurred. The space in which all incident locations are found is divided into smaller areas (sectors) to which ESV resources are allocated. The spatial distribution of incident locations relative to ESVs and hospitals is a key determinant of response times and is important for model validation. *Detail:* ▪ Attribute: Position (in terms of some co-ordinate system)

(Continues on next page)

Table 3.1 (Continued)

Component	Description and Detail
Holding Point	*Description:* A holding point is a position in two-dimensional space where an ESV is located when available. Holding points are chosen for their proximity to areas with a high density of incidents, as a means of keeping response times as low as possible. Ambulances are moved between holding points as demand dictates (see process logic in Fig 3.5) in models making use of dynamic ESV location. *Detail:* ▪ Attribute: Position (in terms of some co-ordinate system)
Sector	*Description:* A sector (also called a drainage area) is a geographically defined area in two-dimensional space. Each sector is associated with one major receiving hospital (the term drainage area refers to the area ”drained” by a hospital, meaning the area from which patients are routed to a specific hospital). Each sector operates more or less independently in terms of response – each sector is resourced with a number of ESVs per shift and those ESVs operate almost exclusively in that sector. *Detail:* ▪ Attribute: Sector Number ▪ Attribute: Defining Boundary

(Continues on next page)

Table 3.1. (Continued)

Component	Description and Detail
Hospital	*Description:* A hospital is a health care facility that can receive patient entities transported to it by ambulances. Each hospital has a geographically defined drainage area and is considered to be the default receiving facility for all patients at incident locations falling within that drainage area, unless otherwise specified. Although patients are received at the EC in a hospital specifically, this level of detail is not necessary in the model as none of the modelling objectives are reliant on it. The only details relevant to the modelling objectives are the spatial relationships between incident locations and hospitals and the handover/make ready delay, which may both affect ESV availability by modulating how long each individual ESV is unavailable for. *Detail:* ▪ Attribute: Position (in terms of some co-ordinate system) ▪ Attribute: Sector (drainage area) ▪ Attribute: Handover/Make Ready Delay

3.4.2.4. Model Process Logic

Model process logic is presented largely in the form of annotated logic flow diagrams, with some accompanying text. Logic of relevant sub-systems of the model is presented in separate diagrams.

Dispatch and Vehicle Allocation Process Logic

Dispatching in this model is considered at a high level of abstraction and the two components of call taking and dispatch are represented as one process, as shown in the upper left corner of Fig 3.2 (on next page). The output of the call taking and dispatch process is an incident which is queued for allocation to an appropriate ESV. The queue is ordered on a first-in-first-out (FIFO) basis, but P1 incidents are always dynamically allocated from this ordering before P2 incidents.

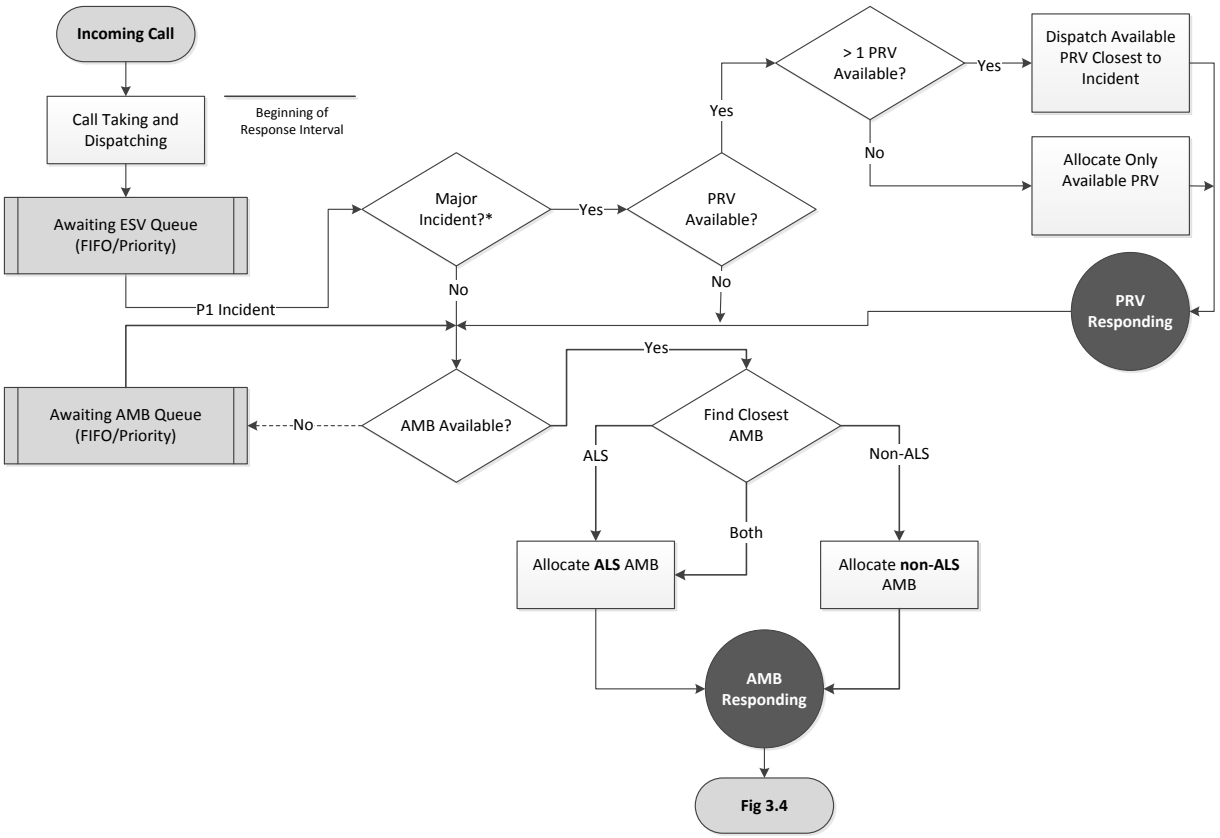


Figure 3.2. Priority 1 Dispatch and Emergency Service Vehicle Logic Flow Diagram
FIFO = First-in-first-out, PRV = Primary Response Vehicle, AMB = Ambulance, ALS = Advanced Life Support, ESV = Emergency Service Vehicle

Process logic for the allocation of ESVs to P1 incidents is shown in Fig 3.2. The first decision point in allocation of an ESV is whether the incident is classified as a major incident. Major incidents are not very clearly defined, but include P1 incidents that would benefit from presence of the highest level of qualified personnel, such as motor vehicle accidents with entrapment or incidents with multiple seriously injured patients. Judgements regarding the classification of a P1 incident as major or not are left to the individual dispatcher and are based largely on their experience. Major incidents involve the allocation of a PRV, followed by one or more ALS ambulances or other ambulance if no ALS ambulance is available. If no PRV is available, the dispatch process continues with only ambulances. Major incidents are relatively rare in the system and consequently the vast majority of dispatching and ESV allocation does not involve PRVs.

As shown in Fig 3.2, allocation of a closest ambulance is dependent on more than one ambulance being available. If no ambulance is available, the incident is queued until at least one ambulance becomes available. The time spent by an incident in such a queue is counted into the total response interval and thus response time. Assuming availability of ambulances, the closest ambulance is determined. If there is only one closest ambulance, this vehicle will be dispatched regardless of

whether it is an ALS or BLS ambulance. If several ALS and BLS ambulances are equidistant to the incident location, an ALS ambulance will always be allocated.

Process logic for the dispatch and allocation of P2 incidents is shown in Fig 3.3. The initial call taking and dispatch process is identical, but there is no consideration of major incidents or involvement of PRVs for any P2 dispatch. The allocation of ambulances is algorithmically similar to that followed for P1 cases, with emphasis on always selecting the closest vehicle. However for P2 cases, if a mix of ALS and BLS ambulances are equidistant to the incident location, a BLS ambulance will always be allocated.

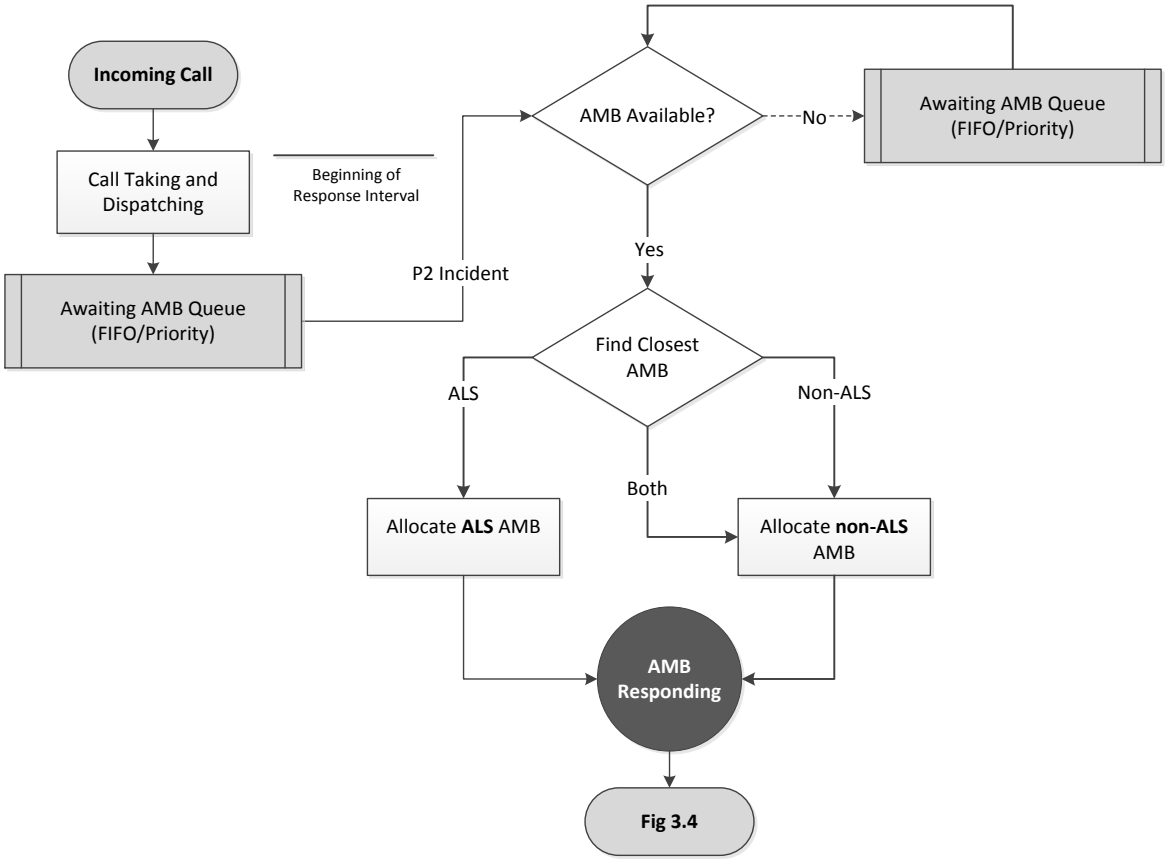


Figure 3.3. Priority 2 Dispatch and Emergency Service Vehicle Allocation Logic Flow Diagram
FIFO = First-in-first-out, AMB = Ambulance, ALS = Advanced Life Support, ESV = Emergency Service Vehicle

Incident Location, Transportation and Hospital Process Logic

Process logic relevant to each incident location, transportation to hospital and activities at hospital are shown in Fig 3.4 (on next page). This diagram does not attempt to differentiate between ESV types (ambulance or PRV) except for the decision to transport to hospital which could apply to an ambulance alone or an ambulance assisted by a PRV. This situation is rare, but would arise if ALS-level treatment is initiated at an incident by personnel who responded in a PRV and transportation

to hospital must be carried out by an ambulance staffed with personnel who cannot continue this level of care.

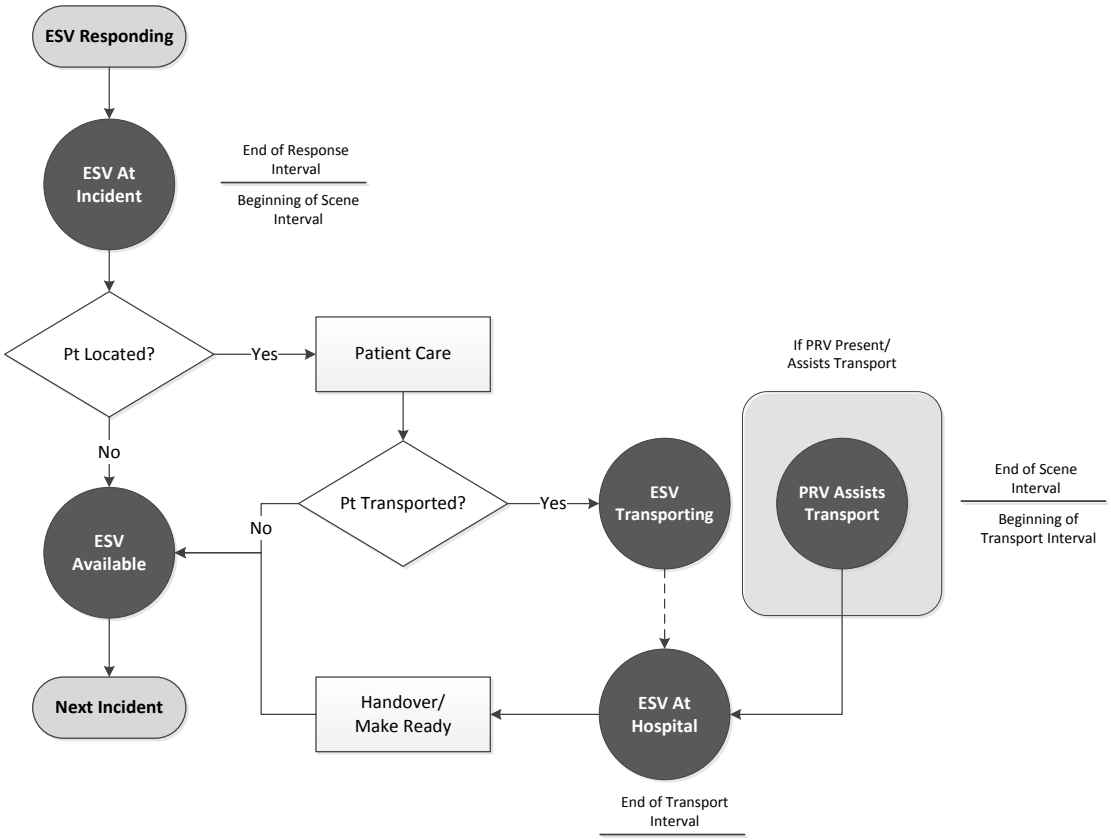


Figure 3.4. Incident Location, Transportation and Hospital Logic Flow Diagram
PRV = Primary Response Vehicle, ESV = Emergency Service Vehicle

The first decision point in Fig 3.4 above relates to whether a patient is located, while the second relates to one or more patients who have been located, assessed and perhaps treated but who are not transported to hospital. In the former case, inability to locate patients may be because they have been removed by another service or by private transport, or because the incident was a hoax and there never was a patient. In the latter case, not transporting to hospital may be the patient's informed decision (either with or against advice) or may be because the patient has been declared dead at the incident scene and will not be transported to a hospital (although they will be removed, but not by EMS resources). From a modelling perspective the only relevant difference is that if a patient is never located there is usually a short delay at the incident scene before the ESV is available again, while in the case of a patient who is not transported there is a longer delay reflecting the fact that some kind of interaction has taken place with the patient. Apart from this difference, the model does not add any further detail to the decisions and processes around whether or not one or more patients are transported to hospital.

The handover and make ready process represents a delay at the hospital and no attempt is made to differentiate between these activities in the model. When this process is completed the ESV state is changed to available indicating that it may again be allocated to an incident following the process logic in Figs 3.2 and 3.3. If there are no outstanding incidents requiring a response the ESV will return to its holding point and await further allocation.

Non-response Ambulance Movement Process Logic

When ambulances are not responding to an incident or otherwise unavailable, they are located strategically at holding points so as to minimise response times. The location of holding points, and the number of ambulances allocated to them, are determined relative to the spatial distribution of incidents – holding points are generally located in or around "hotspots" or areas with a high density of incidents per unit area and time.

If the number of available ambulances at a holding point (referred to here as capacity) decreases to zero (i.e. all ambulances assigned to that holding point are allocated to an incident), a dispatcher will move at least one available ambulance from the closest holding point to the holding point with no capacity, provided there is more than one ambulance available at the closest holding point. When at least one ambulance assigned to the original holding point becomes available again (i.e. capacity ceases to be zero), the replacement ambulance is returned to its original holding point. Any ambulance at or moving between holding points that is available may be allocated to an incident at the dispatcher's discretion, normally based on proximity to the incident. This movement of ambulances is depicted in Fig 3.5 (on the next page).

Non-response movement of ambulances can also occur in a similar way but at a broader level, between sectors. If the capacity of an entire sector falls to zero, one or more available ambulances from a neighbouring sector will be allocated. Once these ambulances are available again they will return to the assigned holding point in their original home sector.

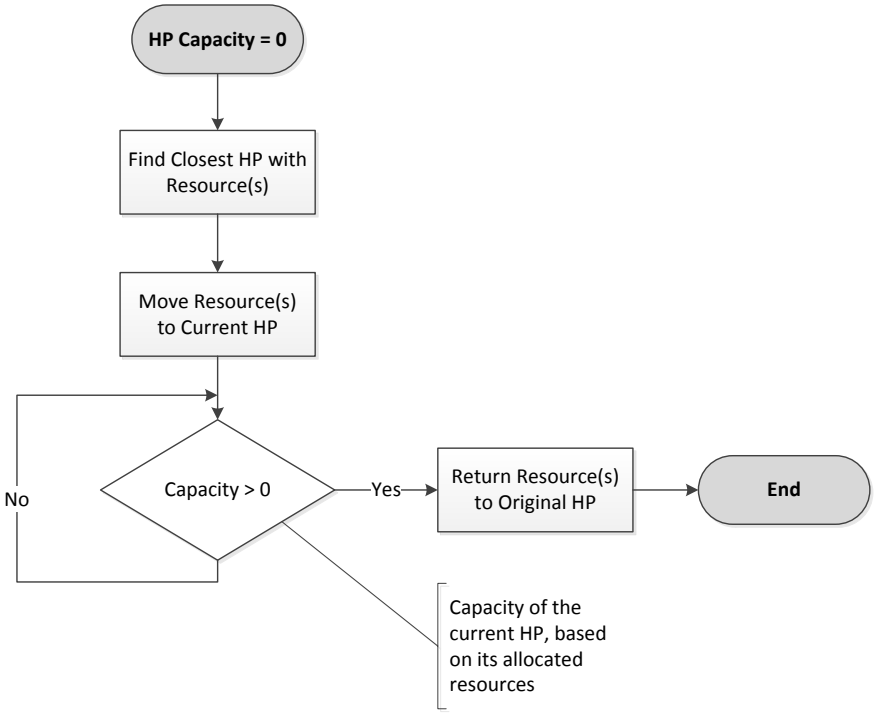


Figure 3.5. Holding Point Emergency Service Vehicle Movement Logic Flow Diagram
HP = Holding Point

Shift Change Process Logic

All ESVs are considered on-shift or available to respond to incidents for an approximate shift duration of 12 hours. At shift changeover, ESV state is changed to off-shift while ESVs are checked by the personnel taking them over (with regard to readiness of the vehicle and medical equipment), after which they return to on-shift state. In order to avoid a situation where ESV capacity in a sector is reduced to zero because of shift changes, these periods are staggered with a variable proportion of the ESVs still available for allocation to incidents. Shift changes are conducted between 06:00 and 07:00 AM/PM.

3.4.2.5. Assumptions and Simplifications

Assumptions and simplifications are an important part of the modelling process. Assumptions are made where there is a lack of clarity about the real system's objects, values or behaviour. This is usually due to a lack of clear policies, rules, documentation or data. Simplifications are intentional attempts to represent elements of the model in a more abstract way than the way in which they occur in reality. This is done in order to make the model easier to understand or implement, or to rationalise the input data requirements. Assumptions and simplifications for the model are given below:

Assumptions

- Shift changes take 30 minutes (duration of off-shift time).
- Staggering of shift changes occurs on a 50:50 basis. In other words 50% of the ESVs in a sector will be off-shift at shift change time while the other 50% are on-shift and available. This is reversed once the first shift change has occurred. ESV selection for the first or second shift change group is on the basis of geographical coverage of the sector area, meaning that at any one time during shift change there will be at least some reasonable coverage of the sector's area.
- The number of ESVs per sector, and per holding point, remains constant for the duration of the simulation.
- ESV speeds during response and non-response states are determined by two components: (i) a constant which is an estimate of average speed for each state and (ii) a weighting factor reflecting estimated traffic congestion by hour of the day. No attempt has been made to model any other effects on ESV speed.
- Mechanical ESV failures are rare and do not need to be modelled.
- Short-term unavailability of ESVs for the purposes of refuelling or re-stocking medical equipment do not need to be modelled.

Simplifications

- No incident involves more than two lying patients.
- A period of average incident occurrence rates is modelled, based on mean daily and hourly incident occurrence rates derived from a continuous 12-month period of data from the real system.
- Only four of six sectors in the real system are modelled, including two sectors that are considered to have a high incident occurrence rate and two that are considered to have a relatively lower incident occurrence rate.
- Each ambulance can transport only one patient entity. After picking up a single patient entity at an incident location, each ambulance will route to the hospital in its sector. Note that a single patient entity could conceptually represent up to two lying patients as most if not all ambulances in the modelled system can carry a maximum of this number of lying patients.
- Once allocated, every ESV immediately becomes mobile to an incident scene. No delays between allocation and becoming mobile are modelled.
- Decisions regarding allocation of ESVs are based upon a set of simple, unequivocal rules such as that the closest ESV is always allocated to an incident, or that the highest priority (P1) incidents

that have been queued the longest are always allocated ESVs first. No attempt is made to model the type of experience-based reasoning and heuristic decision-making that dispatchers sometimes employ in deciding on the allocation of ESVs.

- All patients requiring transportation to hospital are transported to the hospital in their sector (i.e. the sector in which their incident location lies).
- An ESV reaches the incident location and spends some variable time delay at that location for all cases where a patient is not transported to hospital.
- At shift change ESVs do not move to another location for their shift change check or re-stocking, they are simply off-shift for a defined period and then on-shift again. If an ESV is not at its holding point when its state changes to off-shift, it will return to its holding point.
- Hospital closures (due to lack of beds) are rare and do not need to be modelled, given the time frames of the simulation that are of interest.

3.4.2.6. Model Process Logic: Changes for Experimental Factors

Dispatch and Vehicle Allocation Process Logic: Single- and Two-tier Response Model

Process flow diagrams for single- and two-tier response models are shown in Figs 3.6 and 3.7 (on the next two pages). The single-tier process logic for P1 dispatch is very similar to that shown in Fig 3.2, the only exception being the absence of PRV allocation as PRVs do not feature in this kind of model.

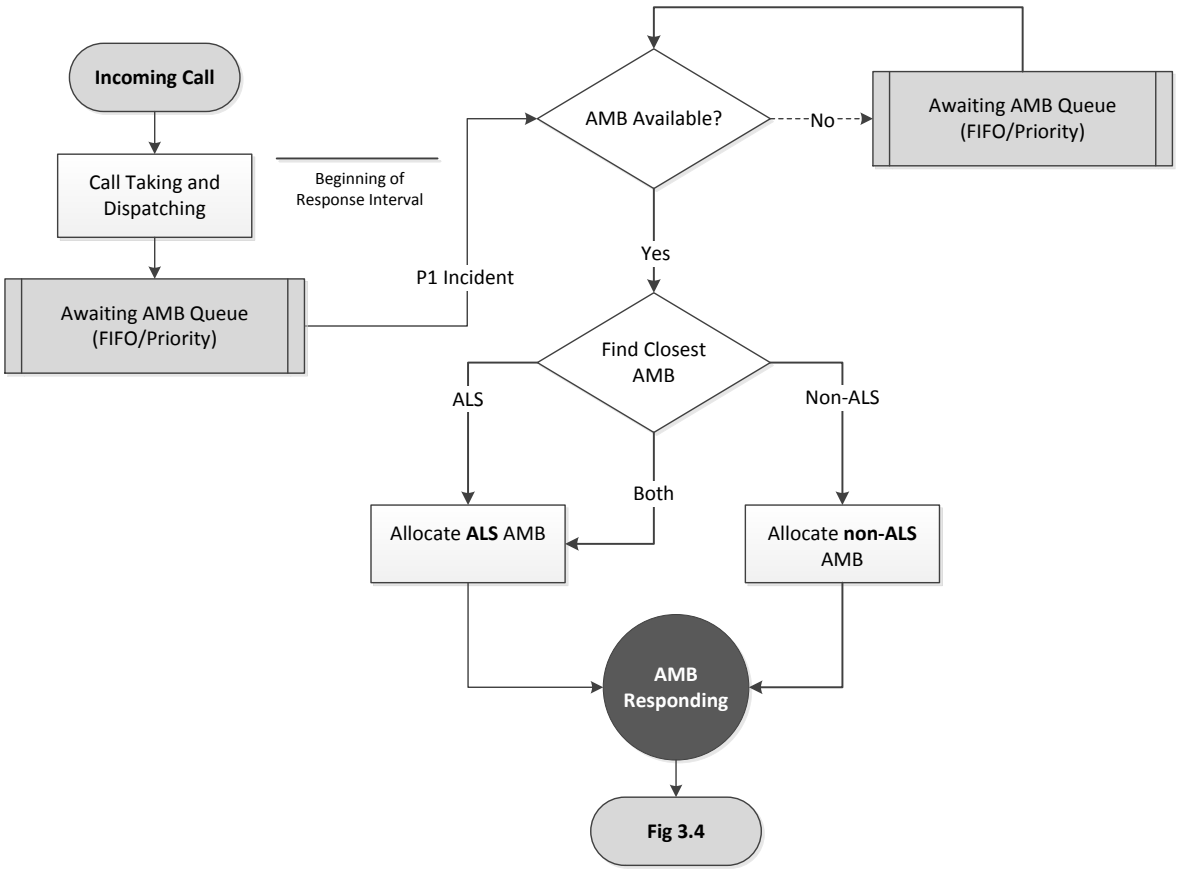


Figure 3.6. Priority 1 Dispatch & Vehicle Allocation Logic Flow Diagram: Single-tier Response
FIFO = First-in-first-out, AMB = Ambulance, ALS = Advanced Life Support, ESV = Emergency Service Vehicle

The two-tier response system process logic shown in Fig 3.7 does include the use and allocation of PRVs, but the rules for this are different in comparison to the process logic shown in Fig 3.2. In this model, PRVs (if available) are dispatched to all P1 incidents rather than just to major incidents. If no PRV is available and more than one ALS ambulance is available then the closest ALS ambulance will be dispatched to the incident, otherwise the only available ALS ambulance will be dispatched. In the event that both a PRV and an ALS ambulance are available, the closer of the two will be dispatched to the incident. If the opposite situation occurs, a non-ALS ambulance will be dispatched, if available. Regardless of the availability of different vehicles, as a rule, a PRV is only dispatched if the ambulance dispatched to the same incident is not an ALS ambulance or the PRV is closer to the incident.

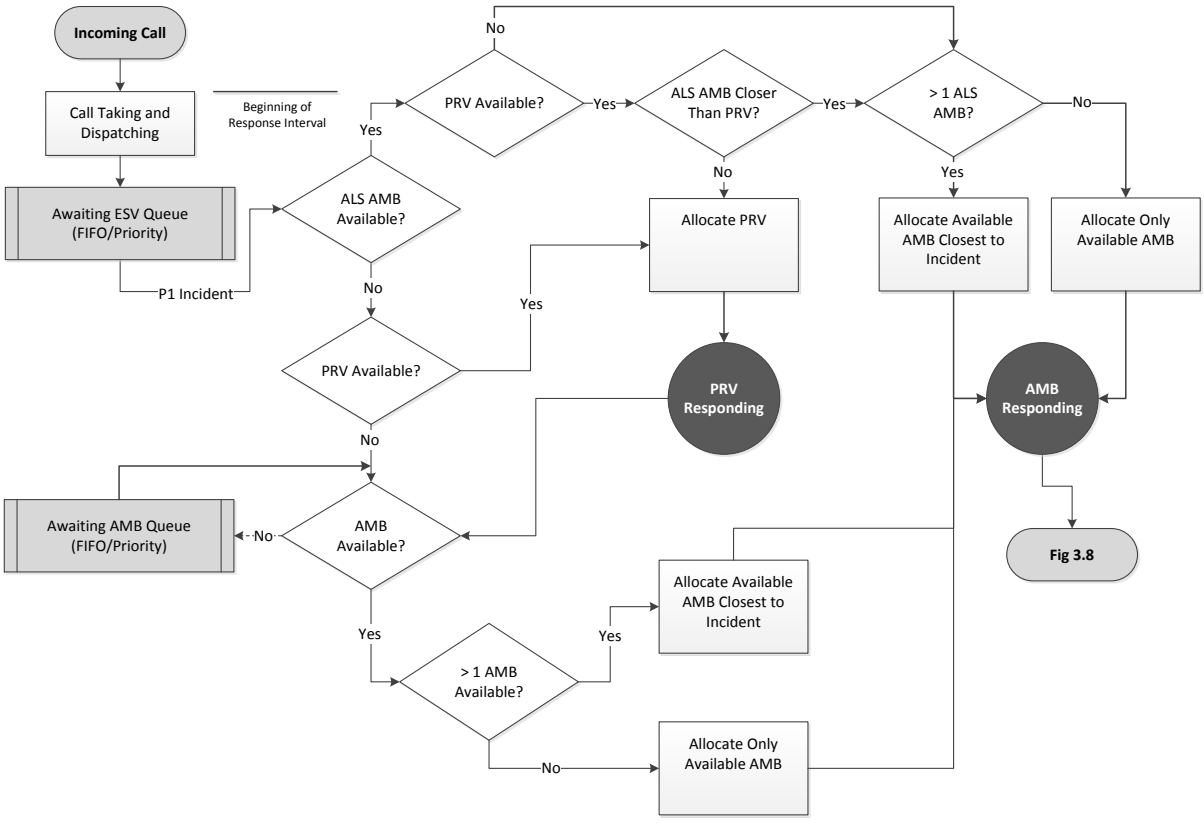


Figure 3.7. Priority 1 Dispatch & Vehicle Allocation Flow Diagram: Two-tier Response
FIFO = First-in-first-out, PRV = Primary Response Vehicle, AMB = Ambulance, ALS = Advanced Life Support, ESV = Emergency Service Vehicle

Process logic for the allocation of ESVs to P2 incidents in the single- and two-tier models is effectively the same as that shown in Fig 3.3.

Incident Location, Transportation and Hospital Process Logic: Single- and Two-tier Response Model

The single-tier response model's process logic related to activities at the incident scene and hospital is similar to that shown in Fig 3.4. The only difference is the involvement of a PRV for assistance in some cases. If this level of care is required at an incident, and this is only recognised after arrival of a non-ALS ambulance (i.e. the case is incorrectly triaged and turns out to be of a more serious nature than first thought), the personnel at the incident scene may request ALS assistance by PRV response. This is shown in Fig 3.8.

Figure 3.8. Incident Location, Transportation & Hospital Logic Flow Diagram: Two-tier Response
PRV = Primary Response Vehicle, AMB = Ambulance, ALS = Advanced Life Support

Important to note in the situation just described, is that if the PRV is located far away from the incident scene and the ambulance personnel are in a position to leave the incident scene before the PRV arrives, they will do so without PRV assistance and transport the patient to the relevant hospital. In such cases, the PRV is cancelled before it arrives at the incident scene and is then available for allocation to another incident, if there is one queued and waiting.

Dispatch and Vehicle Allocation Process Logic: Static and Dynamic Vehicle Location

Static positioning of ESVs, whether in a single- or two-tier response model, effectively means the elimination of all process logic related to positioning of ESVs at holding points and demand-based

movement of ESVs between holding points (as shown in Fig 3.5). Instead, ESVs are stationed at each sector hospital and return to this location when they are available for incident response.

3.4.2.7. Assumptions and Simplifications: Changes for Experimental Factors

Assumptions and simplifications relevant to the baseline model derived from the real EMS system in Cape Town have been listed in 3.4.2.5. In this Section, additional assumptions are made because these relate to the alternative models based upon the two experimental factors (response model and ESV location). Simplifications listed in 3.4.2.5 are also applicable to both of these models.

Table 3.2. Response Model Factor Assumptions

Factor Levels	Assumption
Single-tier Response	▪ There are no PRVs. ▪ The number of ambulances (ALS and non-ALS) is the same as that in the baseline model. ▪ The proportion of ALS ESVs available in the system is the same as that in the baseline.
Two-tier Response	▪ The number of ALS ambulances in each sector is one less than the number in the baseline model, because of the addition of ALS capability in the form of a PRV in each sector. ▪ A PRV is dispatched first to every P1 incident, unless no PRV is available or an ALS ambulance is closer to the incident. ▪ An ambulance is dispatched after a PRV to every P1 incident, unless no PRV is available in which case only an ambulance is dispatched (ALS, if available), or unless an ALS ambulance was closer to the incident. ▪ In some cases (an estimated 10%), a non-ALS ambulance may request ALS assistance from a PRV while at a P2 incident scene. ▪ In some cases (an estimated 70%), a PRV will be available on termination of activities at an incident scene while an ambulance transports the patient to hospital. In other cases (an estimated 30%), a PRV will proceed with an ambulance to hospital.

Table 3.3. Emergency Service Vehicle Location Model Factor Assumptions

Factor Levels	Assumption
Static ESV Location	▪ ESVs are located at the hospital in their sector and no positioning at holding points occurs. ▪ No non-response movement of ESVs occurs, other than return of ESVs to the hospital in their sector when not allocated to an incident.
Dynamic ESV Location	▪ Assumptions are as listed in 3.4.2.5.

3.5. INPUT DATA

3.5.1. Source and Range of Input Data

Input data were obtained from the Western Cape EMS Dispatch Centre's Computer Aided Dispatch (CAD) system (Global Emergency Management Command and Control Centre, version 4.4, Aurecon SA (PTY) LTD, Pretoria, South Africa). The following data fields were obtained, covering the time period between 1 January 2011 and 29 August 2013:

Table 3.4. Computer Aided Dispatch Database Fields

Field	Used for Input Analysis	Description
ir_number	No	Unique record identifier
priority	Yes	Incident priority
latitude	Yes	Incident location, latitude in decimal degrees
longitude	Yes	Incident location, longitude in decimal degrees
suburb	No	Suburb in which the incident occurred
sector	Yes	Sector in which the incident occurred
inc_group	Yes	Type of incident
call_date	Yes	Date/time the incident was opened by a call taker
register_time	No	Date/time the incident was entered into the CAD system for dispatch
dispatch_time	No	Date/time the incident was dispatched
accept_time	No	Date/time the dispatched incident was accepted by a ESV
enroute_time	Yes	Date/time the ESV became mobile to the incident

(Continues on next page)

Table 3.4 (Continued)

Field	Used for Input Analysis	Description
intervene_time	Yes	Date/time the ESV arrived at the incident scene
document_time	No	Date/time of report back to the dispatcher
convoy_time	Yes	Date/time the ESV left the incident scene
patrolfree_time	Yes	Date/time the ESV was available to be allocated to another incident
completion	Yes	The place where the incident was completed (usually a hospital)
exemptindicator	Yes	Flag, indicating whether the incident was exempt, meaning that a patient was either not located or not transported to hospital

ESV = Emergency Service Vehicle, CAD = Computer Aided Dispatch, ESV = Emergency Service Vehicle

This data set was comprised of 825,237 incident records.

3.5.2. Processing and Analysis of Input data

The data referred to above were supplied in plain text format and imported into a SQL Server database (Microsoft SQL Server 2008 R2, Microsoft Corporation, Washington, USA). Text data were transformed into specific data types such as dates or strings using Structured Query Language (SQL) statements during or after the importing process. Only data between 1 January 2012 and 31 December 2012 were used for input analysis as this represented the most recent complete 12-month period.

Only data falling into four of the six sectors supplied were used. The six sectors were those representing drainage areas of Groote Schuur, GF Jooste, Tygerberg, Victoria, New Somerset and Helderberg hospitals, of which only data representing incidents in the Groote Schuur, GF Jooste, Tygerberg and Victoria drainage areas were retained. Incident data flagged as inter-hospital transfers and other agency responses were excluded, leaving only P1 and P2 incidents representing primary responses in the four sectors described above. This data set was comprised of 312,387 incident records.

3.5.3. Data Description

3.5.3.1. Arrival Rate Data

Incident (patient) arrival rate data were obtained by extracting the average number of incidents for each day of the week (Monday – Sunday) and each hour of the day from the 312,387 records, using SQL statements. These data were separated by sector and are shown in Annexures A to D.

3.5.3.2. Incident Priorities

Proportional distribution of incident priorities were based upon SQL statement counts in each priority category (P1 and P2) and were separated by sector. These data are shown in Table 3.5:

Table 3.5. Distribution of Incident Priorities Across Sectors

Sector	P1	P2	Total
GSH (Sector 1)	17,073 (50%)	17,316 (50%)	34,389
GFH (Sector 2)	40,485 (45%)	48,777 (55%)	89,262
TBH (Sector 3)	58,107 (43%)	76,005 (57%)	134,112
VH (Sector 4)	24,687 (45%)	29,937 (55%)	54,624
			312,387

GSH = Groote Schuur Hospital, GFH = GF Jooste Hospital, TBH = Tygerberg Hospital, VH = Victoria Hospital
P1 = Priority 1, P2 = Priority 2

3.5.3.3. Exempt Incidents

Exempt incidents are incidents where a patient is either not located or not transported to hospital for a variety of reasons. The proportion of exempt incidents was based upon SQL statement counts where a filter identifying the exempt flag was used. These data are shown in Table 3.6:

Table 3.6. Distribution of Exempt Incidents Across Sectors

Sector	Exempt		Not Exempt		Total
	P1	P2	P1	P2	
GSH (Sector 1)	5,853 (34%)	5,863 (34%)	11,220 (66%)	11,453 (66%)	34,389
GFH (Sector 2)	12,130 (30%)	15,382 (32%)	28,355 (70%)	33,395 (68%)	89,262
TBH (Sector 3)	19,583 (34%)	29,190 (38%)	38,524 (66%)	46,815 (62%)	134,112
VH (Sector 4)	8,614 (35%)	10,803 (36%)	16,073 (65%)	19,134 (64%)	54,624
					312,387

GSH = Groote Schuur Hospital, GFH = GF Jooste Hospital, TBH = Tygerberg Hospital, VH = Victoria Hospital
P1 = Priority 1, P2 = Priority 2

3.5.3.4. Distributions

A number of statistical distributions for time intervals in the model were derived from the input data set. An input analyser application (Arena Input Analyzer, version 14.5, Rockwell Automation Inc., Milwaukee, USA) was used for this purpose. One of the challenges in deciding on an appropriate distribution fit for a set of input data is the influence that sample size has on statistical hypothesis testing of the given distribution fit (as described in 2.5.3.2 of Chapter 2).

Very large sample sizes will tend to produce a significant result indicating that, at the chosen significance level, the null hypothesis of the sample distribution not being the same as the hypothesised distribution should be rejected. This may occur even if there is a seemingly appropriate fit demonstrated by visual assessment of the sample distribution.(1,63) A number of approaches may be used in such circumstances to make a decision about the best fitting distribution. The magnitude of the goodness-of-fit p-value can be considered rather than just considering it as a threshold accept/reject indicator. In addition, the shape of the distribution can be assessed and compared to known distribution shapes, or a smaller random sample of data can be selected from the original sample and goodness-of-fit tests applied to it.(1,63,70) The magnitude of the square error associated with each fitted distribution may also be considered. Results of the model-fitting process are shown in Table 3.7:

Table 3.7. Input Data: Fitted Probability Distributions

Interval	Distribution/Parameters	Basis of Fit Decision
P1 Dispatch Interval *(dispatch_time - call_date)*	97 * Beta(1.42,20.3)	All goodness-of-fit p-values were < 0.05. Visual confirmation/smallest square error were used as criteria for distribution fitting
P2 Dispatch Interval *(dispatch_time - call_date)*	423 * Beta(0.589, 4.5)	
P1 Scene Time Non-Exempt *(convoy_time - intervention_time)*	Gamma(11.4, 3.53)	
P2 Scene Time Non-exempt *(convoy_time - intervention_time)*	Gamma(11.4, 3.53)	
P1 Scene Time Exempt *(convoy_time - intervention_time)*	Exponential(22.7)	
P2 Scene Time Exempt *(convoy_time - intervention_time)*	Exponential(15.4)	

(Continues on next page)

Table 3.7 (Continued)

Interval	Distribution/Parameters	Basis of Fit Decision
Hospital Delay Sector 1	Exponential(25.52)	Knowledge of process and mean values (the source of mean values is described in 3.5.4.3)
Hospital Delay Sector 2	Exponential(21.49)	
Hospital Delay Sector 3	Exponential(26.99)	
Hospital Delay Sector 4	Exponential(21.95)	

P1 = Priority 1, P2 = Priority 2, Sample of Distribution Histograms Shown in Annexure L

3.5.4. Other Data and Sources

The input data described above were extracted directly from the dispatch centre's CAD system. Other data relevant to the conceptual model were not stored in the CAD system and were obtained through interviews with dispatch centre staff. These are discussed below in 3.5.4.1 - 3.5.4.3.

3.5.4.1. Emergency Service Vehicle Numbers

ESV numbers allocated to each sector per shift may vary over time. This is based primarily on factors influencing the number of ESVs and staff available for a given shift. As noted under model simplifications in 3.4.2.5, it was not considered necessary to model this short-term variation is ESV numbers but rather to keep ESV numbers constant and a reflection of what could be considered typical for a given shift, as shown in Table 3.8:

Table 3.8. Typical Emergency Service Vehicle Numbers per Sector

Sector	Non- ALS Ambulances	ALS Ambulances	Total Ambulances	PRVs
GSH (Sector 1)	9	3	12	1
GFH (Sector 2)	10	3	13	1
TBH (Sector 3)	12	3	15	1
VH (Sector 4)	7	2	9	1
	38	11	49	4

GSH = Groote Schuur Hospital, GFH = GF Jooste Hospital, TBH = Tygerberg Hospital, VH = Victoria Hospital, PRV = Primary Response Vehicle

3.5.4.2. Holding Point Numbers, Locations and ESV Allocations

Although holding point locations are generally constant over short periods of time, numbers of ESVs allocated to each holding point can vary depending on the total number of ESVs available per shift. The data presented below are an indication of a typical allocation of ESVs to holding points, given the ESV numbers per sector in Table 3.8. Important to note is that ESVs may move between holding points depending on demand, but remain allocated to specific holding points as shown in Table 3.9.

This means that a given ESV will ultimately return to its allocated holding point in the event that it is moved to a different holding point because of an increase in demand in a specific area.

Table 3.9. Typical Allocation of Emergency Service Vehicles to Holding Points

Sector	Holding Point	Ambulances	PRVs
1 (GSH)	GSH [HP1]	4 + 1 ALS	
	Pinelands [HP2]	1 ALS	1
	Heideveld [HP3]	1	
	Hanover Park [HP4]	1 ALS	
	Vanguard [HP5]	3	
	Gatesville [HP6]	1	
		12	1
2 (GFH)	Sector 2 EMS Base [HP1]	1 + 1 ALS	1
	Mitchell's Plain CHC [HP2]	2 + 1 ALS	
	GFH [HP3]	2	
	Gugulethu CHC [HP4]	2 + 1 ALS	
	Phillipi [HP5]	2	
	Strandfontein [HP6]	1	
		13	1
3 (TBH)	Delft [HP1]	3	
	Eerste Rivier [HP2]	1 ALS	
	Kuils River [HP3]	1	
	Kraaifontein [HP4]	1	
	Bellville [HP5]	3	
	Goodwood [HP6]	1	
	Elsies River [HP7]	2 + 1 ALS	
	Bishop Lavis [HP8]	1	
	Durbanville [HP9]	1 ALS	
	TBH [HP10]	0	1
		15	1
4 (VH)	VH [HP1]	1	1
	Wynberg [HP2]	1	
	Retreat [HP3]	2 + 1 ALS	
	Ottery [HP4]	1 + 1 ALS	
	Hout Bay [HP5]	1	
	False Bay (Fish Hoek) [HP6]	1	
		9	1

HP = Holding Point, GSH = Groote Schuur Hospital, GFH = GF Jooste Hospital, TBH = Tygerberg Hospital, VH = Victoria Hospital, PRV = Primary Response Vehicle, ALS = Advanced Life Support

3.5.4.3. Hospital Waiting Times

Data on how long ambulance personnel are required to wait to complete handover at each hospital are not routinely recorded in the CAD system which served as the source for other data discussed above. Shortly before the time of data collection for this study, personnel at the Western Cape EMS Dispatch Centre had initiated a survey of these times at major hospitals, including those associated with the four sectors included in this model. As a result of this process average hospital waiting times were obtained - unfortunately no indication of dispersion or distribution in these data could be obtained.

The data were recorded over several months and the most recent average hospital waiting times (from September 2013) were used in the model. These are: (i) Groote Schuur Hospital (Sector 1) [25.52 minutes], (ii) GF Jooste Hospital (Sector 2) [21.49 minutes], (iii) Tygerberg Hospital (Sector 3) [26.99 minutes] and (iv) Victoria Hospital (Sector 4) [21.95 minutes]. The distributions used for these waiting times are given in Table 3.7.

3.6. MODEL TRANSLATION AND THE COMPUTER MODEL

3.6.1. Simulation Software

The conceptual model described above was translated into a software representation using an object-orientated simulation application called Simio (Simio Design Edition, version 6.97, Simio LLC, Pennsylvania, USA). Simio approaches simulation from the object-orientated paradigm, using a simulation framework based on intelligent objects. A simulation model is composed of intelligent objects which can either be fixed or move through the model interacting with other objects according to predefined or custom-developed behaviour.

Importantly, model development in Simio does not require programming to leverage the advantages of object-orientation. Unlike its predecessor object-orientated simulation frameworks, Simio is built upon a rich graphical user interface and animation engine making model development easier, more intuitive, quicker and less error-prone. Simio provides a wide array of supporting features for modelling and experimentation including sophisticated debugging tools, data analysis features, experiment design and control features and an optimisation add-in.(117)

Simio objects are created graphically and can have their basic behaviour configured for specific purposes. Following the object-orientated paradigm, all objects are derived from a base class with certain fundamental properties, states and processes. Through a process of sub-classing, Simio

exposes a standard library of objects representing a mix of fixed and moveable objects such as nodes and transporters. The behaviour of each of these standard library objects can be extended and customised by further sub-classing and by the addition of process logic. Process logic is implemented in a set of process steps that can be used by the modeller in combination to bring about virtually any kind of intelligent behaviour that is desired.(117) An example of process steps in Simio's modelling interface is shown in Annexure M.

Simio was chosen as the modelling platform for this simulation because of its flexibility as described above, but particularly because of its support for models involving vehicles and transportation. This is accomplished through a standard vehicle object which can be configured to select entities according to an easily-customisable set of rules, pick up and transport any number of entities and drop off carried entities at specified destinations. Simio vehicles can also be configured to travel along pathways or through two-dimensional space at any speed and in any direction, and can be customised to behave in more complex ways through the addition of process logic. The two-dimensional space in which vehicles operate can be animated, allowing their behaviour to be observed over a range of system conditions and states, which is an important part of the model verification process. The following sections will describe how each part of the conceptual model was implemented using Simio.

3.6.2. Model Objects and Behaviour

3.6.2.1. Incident Locations and Sectors

Representation of Incident Locations

Simio provides a two-dimensional (or three-dimensional, depending on the view chosen) space for visual modelling of model objects, known as the facility window. This space is divided into cells by a grid, the relative dimensions of which are controlled by zooming into or out of the window. Incident locations were modelled using transfer node objects (referred to here as "incident nodes") in Simio, which represent a point in the facility widow which offers processing logic controlling entity movement into and out of it, and other related behaviour. The model input data referred to in 3.5 contained geographic co-ordinates for each incident, and the incident nodes referred to here are the simplified computer model equivalent of the points described by the input data co-ordinates.

Clustering and Spatial Distribution of Incident Locations

Given the large number of incidents, and thus incident co-ordinates in the input data, it was necessary to find a way to simplify the representation of incident locations in the computer model,

as it was not feasible to create an incident node in the computer model for every set of incident co-ordinates in the input data. The approach taken was to divide the geographic area of the four sectors in the real system into a grid of 2 km x 2 km cells and to cluster the incidents occurring within each of these cells. This grid could then be replicated in the Simio facility window and an incident node placed in the middle of each cell representing the clustered incidents from the real system.

Clustering of incidents from the input data was accomplished by first plotting all of the incident co-ordinates as points on a map of the City of Cape Town and surrounding area using Geographic Information System (GIS) software (ArcGIS, version 10.2, Environmental Systems Research Institute, California, USA). A grid overlay was created to cover the extent of the plotted points using the fishnet tool in ArcGIS. The map layer containing incident points was then joined with the grid and an incident point count per cell of the grid was performed.

The first incident point count yielded data in some cells that appeared erroneous. For example, certain grid cells showed counts of several hundred (or more) incident points but inspection of points on the map in the corresponding cell indicated a much lower count. After discussion with a CAD system expert, it was explained that this anomaly was due to behaviour of the CAD system in geocoding of address information. Specifically, in cases where an incident address could not be found by a dispatcher using the CAD system graphical user interface, the system allowed manual entry of the incident address but documented the incident co-ordinates as being the centroid of the suburb selected by the dispatcher. This accounted for the mismatch between incident point counts and observed incident points in some cells.

Further analysis of the 312,387 incident records for 2012 showed that 18,061 unique combinations of latitude and longitude existed in the data set, representing some 6% of the total number of incidents. In some cases, duplicated co-ordinates most likely did represent incidents at the same address or location. However many of these, particularly those with a large number of duplicated co-ordinates simply represented centroid points of suburbs in which incidents without translatable addresses had been mapped.

In order to remove the effect of duplicated incident points, these 18,061 unique points were retained and all others were deleted. This resulted in all geocoded addresses being retained as incident points, plus at most one of each suburb centroid-mapped incident point. These data are shown in Annexure F, where hot spot analysis was performed in ArcGIS. Red areas are those where

clusters of significantly high incident numbers are surrounded by similar clusters, while blue areas are those where clusters of significantly low incident numbers are surrounded by similar clusters. Beige areas do not represent significant clusters of either high or low incident numbers. The distribution of incidents shown in Annexure F was verified as being approximately correct, according to knowledge of the system by experts experienced in dispatch operations and planning.

Taking the total number of incidents per sector and dividing the grid cell counts into this (for each sector) produced a proportional weighting for each grid cell, indicating a relative distribution of incidents per grid cell. These data were used primarily for the allocation of patient entities to incident nodes in the computer model, so that this distribution was proportionally similar to that from the input data, as shown in Annexure F. In simple terms, this ensured that "busy" areas in the real system (with a high density of incidents per unit area) were replicated as "busy" areas in the facility window space, and *vice versa* for "quiet" areas.

Representation of Sectors

The spatial extent and configuration of sectors in the computer model was based upon the equivalent extent and configuration identified in the GIS map. The 2 km x 2 km GIS grid referred to above mapped directly to the Simio facility window grid and this allowed each grid cell in each sector to be mapped to an equivalent cell in the facility window. Thus each incident node in the facility window was placed into one of the four sectors, depending on where the centroid of the equivalent cell in the GIS map occurred. A simplified outline of each sector was drawn around each sector's grid cells in the facility window with a polygon drawing object. This was done in order to make identification of each sector easier during observation of animated behaviour of the model, for testing and verification purposes. A GIS map of all sectors is shown in Annexure G. Individual sector maps and the equivalent modelled representations are shown in Annexures H - K, depicting the locations of hospitals and holding points.

Incident nodes in Annexures H - K are dark red diamond-shaped objects and have been placed in the middle of each 2 km x 2 km grid cell (the visible blocks are 1 km x 1 km, with four of them creating a 2 km x 2 km cell). Shaded polygons are shown giving a simplified outline of sectors (with the exception of Sector 3) and labels identify hospitals and ESV holding points. Incident nodes placed outside of sector boundaries were left in position for orientation during development of the model, but were not the destination of any patient entities during execution of the simulation (this is explained in 3.6.2.2 about patient entity routing to incident nodes).

3.6.2.2. Patient Entities

Patient Entities and Entity Creation

Patient entities were modelled using the default *ModelEntity* object in Simio. Two different patient entities were used, representing P1 and P2 patients. Several state variables were added to the existing set of patient entity states, including states reflecting whether the patient is exempt (i.e. not transported to hospital), whether the PRV should assist transportation to hospital (two-tier response model) and whether assistance is required from a PRV (two-tier response model). A number of other state variables reflected the patient priority, the entity' sector and a number of times, such as when the entity entered and left dispatch, which were used later in in the simulation.

Although in the conceptual model (as in any real system) patients exist at an incident scene when the event that causes their medical emergency (e.g. a motor vehicle accident or some acute pathological event) occurs, this could not be modelled in Simio. An approach was thus followed where patient entities were created, configured, subjected to a delay simulating the dispatch process and then routed to an incident location. Once at the incident location, each patient entity would initiate a sequence of events resulting in the movement of a vehicle (ambulance) to the incident and subsequent transport of the patient entity to a hospital. Important to note is that movement of patient entities between the source object, transfer node and incident nodes did not incur a time penalty.

Patient entities were created by a source object which allowed the arrival rate of these entities to be specified. Arrival rate data were set in a rate table, using the data shown in Annexures A to D. Rate tables is Simio are used to simulate time-varying arrival rates following a non-stationary Poisson process, with piecewise-constant arrival rates specified in per hour units in the table. After entities were created, but before they exited the source object, each of the relevant state variables was assigned a value. These values were obtained from another data table, following values or distributions specified in that table which allowed the desired proportional distributions to be followed (as indicated in Table 3.5 for priorities, as an example).

The Dispatch Delay and Transfer of Patient Entities to Incidents

A patient entity source as described above was created for each of the four modelled sectors and was responsible for creating patient entities for only that sector, in accordance with arrival rate and input data distributions for that sector. Once created, each patient entity was routed to a transfer

node object. This movement was accomplished in zero time. On entry into the transfer node object, each patient entity underwent a delay simulating the dispatch process.

Available CAD data allowed calculation of the interval from the beginning of call taking until the dispatch of each ESV, which consists of two parts - the time taken for call taking and preparation for dispatch, and the time taken for a suitable ESV to become available in order to be dispatched. The latter interval is referred to in Fig 1.1 as the hand-off delay; that is the delay in handing-off an incident that is ready to be dispatched to a suitable ESV. Depending on ESV availability at any given time, the hand-off delay may vary in duration. Unfortunately, the hand-off delay as a specific interval was not captured in the CAD data which only reflected the total interval between the beginning of call taking until the actual dispatch of each ESV. The model, however, by simulating the interaction between incident arrivals and ESVs did incorporate the hand-off delay in a realistic way as fluctuations in ESV availability meant that an ESV was not always immediately available for dispatch at times when ESV availability was low.

During the model validation process, output data indicated that total P1 response time was on average 2.30 minutes longer than that observed in the system output data, despite reasonable values for P1 travel response time. The only possible reason for this effect was that the hand-off delay had effectively been represented twice in the model - once in the dispatch delay as derived from CAD data and once in the behaviour of the model, based on variations in ESV availability. For this reason, the total P1 dispatch delay obtained from CAD data was shortened by a factor producing total response times in accordance with those obtained from system data. In this way, the P1 hand-off delay of 2.30 minutes on average was produced only by the model's behaviour based on ESV availability and was excluded from the dispatch delay derived from CAD data. The system time at the start and end of the dispatch delay was recorded in two state variables for later use in measuring the associated response time and other time intervals.

After the dispatch delay, each patient entity was transferred to an incident node. A transfer in this sense means that the entity was moved from the transfer node (its current location) to an incident node in zero time. A decision regarding which incident node to transfer a given patient entity to was made on the basis of random allocation, however spatial distribution of real incident locations in the input data was factored into this. A reference to each incident node in each sector was stored in a data table from which a patient entity's transfer destination was determined. In the same table, for each incident node reference, a number was stored indicating that node's proportional weighting in

terms of the overall spatial distribution of incidents in a sector (as described in 3.6.2.1). Random allocation of patient entity routing destinations per sector was executed over the duration of a simulation run by Simio in accordance with the proportional weightings referred to above, ensuring that the spatial distribution of incidents in each sector's incident node set was similar to that identified in the input data.

Arrival of Patient Entities at Incident Nodes: Process Logic

Arrival of a patient entity at an incident node resulted in a sequence of processes occurring, partly handled automatically by Simio and partly modified with the addition of process logic. Vehicle allocation to, pickup and transporting of entities in Simio follows a sophisticated sequence of events and processes that are part of the innate intelligence of these objects. In summary form, the arrival of an entity at an incident node results in a visit request being sent to a set of vehicles (ambulances), the best of which will be allocated to the entity as its transporter. "Best" is determined by configuration options for the set of vehicles (discussed further in 3.6.2.3). The entity waiting for pickup then resides in a ride station which is part of the incident node until the reserved vehicle arrives at the node for the pickup. This basic behaviour was used without modification.

In addition to the above, each incident node has a *node_entered* event triggered when any object enters it. Additional processes can be defined and set to execute whenever this event is triggered allowing process logic to be defined for any desired purpose. This was done in the current model in order to modulate response behaviour when a P1 patient entity entered an incident node, requiring the dispatch of a PRV (which was a process independent of that outlined above for the pickup of a patient entity by an ambulance).

3.6.2.3. Emergency Service Vehicles

Vehicles in Simio

Vehicles in Simio are a specialised type of object with a substantial amount of built-in behaviour. Many aspects of this behaviour can be customised simply by setting configuration options, but as with all other objects, more detailed control of vehicles is possible with the addition of custom process logic. Although a single vehicle object is configured for its specific purpose at design time, at run time this single object serves as a template for a dynamic population of vehicles with the same properties and behaviour.

In Simio's interactive mode, where it is possible to view an animation of the simulation space, vehicles can be observed moving and interacting with other objects such as nodes. Two types of vehicle movement are possible: (i) fixed route, where vehicles move along paths connecting nodes according to a fixed sequence and (ii) on demand, meaning that vehicle movement can occur at any time, as required. In addition, vehicles can be configured to move only on a network of paths or in what is referred to as "free space" – in a straight line between any two points in the two-dimensional facility-window.

Vehicles in the current model were configured to move on demand and in free space in order to facilitate the most realistic type of movement possible. This allowed vehicles to be diverted from one free space trajectory to another in the event that they were dispatched while moving back to a holding point or if they were cancelled while on their way to an incident node (this applied to PRVs, see below) and diverted to a different node.

Transporters vs. Non-transporters: Ambulances and Primary Response Vehicles

Any vehicle can be configured to fulfil the role of a transporter or non-transporter, meaning that it has the capacity to transport one or more entities or not. A significant amount of built-in logic allows a transporter to be selected (on specified criteria) for a transport task, moved to the correct node, loaded with an appropriate entity , routed with the loaded entity to an appropriate destination and unloaded at that destination. Ambulances were modelled as transporters with a capacity to carry one patient entity whose destination was set to the relevant sector's hospital node.

A patient entity at an incident node in a given sector was configured to select an ambulance from a list for that specific sector. Each sector was associated with a list containing references to a group of ALS and non-ALS ambulances, thus only those ambulances would be considered for the transportation of patient entities in that specific sector, modelling the approach that ambulances were allocated for work primarily only in one sector. As discussed in the conceptual model, under non-response movement of ambulances, situations may arise where there are no vehicles available in a sector. In such cases an ambulance may be "borrowed" from a neighbouring sector in order to attend to an incident and then "returned" to its original sector on completion. This behaviour was modelled by including a reference to neighbouring ambulances in each sectors ambulance list, but placing this as the last item in the list. With the list selection option set to selection in the preferred order, neighbouring ambulances would only be allocated to incidents in a given sector if all of the ambulances higher up in the list had been utilised and were not available for allocation.

PRVs were modelled as non-transport vehicles in keeping with the mode of operation in the system being modelled. This required that much more of the process logic related to selection and movement of PRVs had to be customised, as the built-in logic for transporters did not apply. Following Simio's paradigm for object resource allocation, wherever a PRV response was required, the relevant patient entity seized capacity of the appropriate PRV which was configured to move to the incident node containing the patient. A seized object is said to be "owned" by the object which has seized it, preventing any other object from acquiring its capacity until the owning object has released it. PRVs were also selected from a list for each sector when a seize operation occurred, and each PRV was active in only one sector. Because of the smaller numbers of PRVs than ambulances per sector, PRVs were not modelled to assist with response beyond their allocated sectors like ambulances.

Ambulance and Primary Response Vehicle Selection Rules

Simio exposes a variety of configuration options for the selection of vehicles, whether they are acting as transporters or non-transporters. Some of these options are set under vehicle configuration, while some are set under configuration options for incident nodes. The following rules for vehicle selection were applied (which are also referred to in Figs 3.2, 3.3, 3.6 and 3.7:

- *Ambulances:* A first-in-first-out ranking rule was used to order patient entity transport requests. In addition, a dynamic selection rule was specified, that requests from P1 patient entities would always be selected first from this ranking. The closest ambulance (as determined by Simio, using a minimising straight-line distance measurement method) was reserved after selection of the most appropriate patient entity. Other selection rules related to ALS and non-ALS ambulances were implemented following the logic of Figs 3.2 and 3.3, by configuration of incident node vehicle selection goals and expressions.
- *PRVs:* A similar first-in-first-out ordering of queued seize requests for each PRV was used, although there was no priority-based dynamic selection from this ordering because only P1 patient entities were eligible to seize capacity of a PRV. As with ambulances above, the closest PRV was always allocated.

One other behaviour of relevance to the selection of PRVs was modelled, to comply with a common sense rule for dispatch of these vehicles. On release, after having been seized by a patient entity at some point, each PRV searched through its queued seize requests (in what is referred to as an allocation queue) and any patient entity waiting to seize a PRV that had already left its incident

location for a hospital was removed from the queue. This models the situation where a dispatcher has scheduled a PRV to respond to an incident when it has completed its current response, but then cancels that future response when the patient has left the incident scene in an ambulance. Clearly, no further action from a PRV is needed when the patient is already *en route* to a nearby hospital.

Behaviour of Vehicles: Incident Nodes and Hospitals

Both ambulance and PRV behaviour after entering an incident node was determined by the addition of custom process logic. This process logic was effectively an implementation of Figs 3.4 and 3.8 using Simio's process steps.

In summary, ambulances underwent a delay at each incident node to simulate on-scene activities, following a distribution as shown in Table 3.7. After this delay, at which time the patient entity had been loaded, the ambulance travelled to the receiving hospital for its sector. At the receiving hospital each ambulance underwent another delay simulating patient handover and making the ambulance ready for the next response. This delay was based upon input data as presented in Table 3.7.

PRVs, when dispatched, were delayed on scene for the same duration as the ambulance at the same incident node. On departure of the ambulance from the incident node, the PRV was either released by the relevant patient entity and was available for another incident response, or travelled with the ambulance to the sector receiving hospital if this was required (i.e. if PRV assistance of ambulance transport was specified). In such cases, the PRV was delayed at the hospital for the same time as the ambulance, simulating patient handover and make ready activities as above.

Exempt Incidents

In the case of exempt incidents (those where a patient is not located or transported to hospital) both ambulances and PRVs were seized using custom process logic and moved to the incident node to simulate the response. This approach was followed as it was easier to control than allowing the built-in ambulance reservation and loading processes described above to occur, and then attempting to stop the ambulance transporting the patient entity to a hospital or remove the patient entity from the ambulances ride station (the location in a transporter object where a loaded entity is located for transportation).

Delay times for exempt incidents were typically shorter than in the case of non-exempt incidents as shown in Table 3.7. After this delay both ambulance and PRV, if present, were released and became available for allocation to the next patient entity wanting to seize their capacity if there was one.

Calling for Advanced Life Support Assistance

In the two-tier response model implementation, a non-ALS ambulance sometimes requested ALS assistance from a PRV while at the incident scene. This was modelled as a short delay after the non-ALS ambulance entered the incident node, followed by the patient entity seizing capacity of the closest PRV and the PRV moving to the relevant incident node. This was followed by a second longer delay by the ambulance at the incident node, simulating patient care activities while awaiting arrival of the PRV.

In such cases, additional process logic was used to ensure that if the PRV had not yet arrived at the incident node by the time the non-ALS ambulance had reached the end of its delay period (i.e. that the ambulance personnel had done what they could at the incident scene and had loaded the patient into the ambulance and were ready to initiate transport to hospital), the PRV was released when the ambulance left the incident node on the way to its receiving hospital. This simulated the cancelling of a responding PRV when the ambulance was ready to leave the incident.

Modelling of Travel Times

Simulation travel time was determined by the straight-line distance between two nodes divided by vehicle average speed. The straight-line distance between nodes was in turn determined by spatial relationships between incident nodes, hospitals and holding points (for models using dynamic vehicle location).

No system data on vehicle average speeds were available and thus all average speed values in the model are estimates, and in many cases were determined by manipulation during model validation. All vehicle average speeds were determined by a base constant and a traffic congestion coefficient. The traffic congestion coefficient was a number ranging between 0.7 and 1.0 estimated to represent the approximate degree to which vehicle speeds could be slowed at certain times of the day due to traffic congestion. Values for vehicle speeds and traffic congestion coefficients by time of day are shown below in Table 3.10.

Table 3.10. Emergency Service Vehicle Average Speed Values and Traffic Congestion Coefficients

Category	Average Speed (km/h)
All Non-response/Non-transport Travel	16.7
P1 Ambulance/PRV Response Travel	30.1
P2 Ambulance Response Travel	16.7
P1 Transport Travel	25.5
P2 Transport Travel	27.3
Time of Day	**Traffic Congestion Coefficient**
06:00 – 07:59	0.95
08:00 – 09:59	0.70
10:00 – 14:59	1.00
15:00 – 15:59	0.98
16:00 – 16:59	0.90
17:00 – 17:59	0.80
18:00 – 18:59	0.90
19:00 – 19:59	0.95
20:00 – 05:59	1.00

P1 = Priority 1, P2 = Priority 2

Traffic congestion coefficients were placed in a lookup table. Each ESV speed setting in the model's process logic included a reference to this lookup table based on a time of day (hour of day) lookup (x) value. Based on this input to the lookup table the relevant traffic congestion coefficient (y value) was returned and multiplied with the ESV speed setting to produce a traffic congestion-weighted final speed. The rate table is designed to return a value based on linear interpolation if the x lookup value lies between values contained in the table.(118)

The P1 and P2 transport average speeds shown above may seem relatively high, however this compensates for the model simplification (as listed in 3.4.2.7) that all patients were transported to a single hospital per sector. In the real system some patients may be transported to other hospitals or clinics closer to the incident scene, thus requiring on average faster transport speeds in the model to offset this effect.

3.6.2.4. Documenting Response Intervals and Other Times

Accurate recording of various sub-sections of the response interval was central to the objectives of this study from an experimental perspective, but also played an important part in verification and

validation of the model. The approach taken for the recording of time points (from which time intervals were calculated) was to save these in state variables attached to patient entity and vehicle objects which were set with the system time at various points as each object progressed through the simulation. Calculated time intervals in Simio are saved to a tally statistic as each simulation replication runs and mean, maximum and minimum values for each tally statistic are reported automatically at the end of the run. The same approach is used when multiple replications are executed and in this case confidence intervals calculated across all replications are automatically compiled and reported. In addition, it is possible to write calculated time intervals to a text file as the simulation runs – an approach that was used during experimental comparison of different models in this study.

The following time intervals were calculated (all in minutes):

Table 3.11. Response Intervals

Interval/Description	Start	End
Dispatch Time *Time taken for the dispatch process to be completed*	Beginning of dispatch delay (in patient entity source transfer node)	End of dispatch delay (in patient entity source transfer node)
Response Time (Travel) *Time taken for the ESV to travel to the incident node, once dispatched*	Allocation/seize of ambulance or PRV	Entry into incident node
Response Time (Total) *Time taken for the ESV to reach the incident node, from when the dispatch process starts*	Beginning of dispatch delay (in patient entity source transfer node)	Entry into incident node
Scene Time *Time spent at the incident node*	Entry into incident node	Exit from incident node
Transport Time *Time taken to travel between the incident node and the receiving hospital*	Exit from incident node	Entry into hospital node

(Continues on next page)

Table 3.11 (Continued)

Interval/Description	Start	End
Hospital Time *Time taken to hand patient over at the receiving hospital and prepare ESV for the next incident*	Entry into hospital node	Exit from hospital node

ESV = Emergency Service Vehicle, PRV = Primary Response Vehicle

In addition to response interval times as described above, ESV availability (specifically ambulance availability) during each simulation run was also measured. This was done by writing the remaining capacity (i.e. ESV availability, defined as available ESVs divided by total ESVs) of a sector's ESV group to a text file every time a ESV was allocated to an incident. The mean availability over a simulation run was calculated from these data, and pooled to give an indication of the system's ESV availability.

3.6.2.5. Holding Points

Holding Points as Nodes

Holding points were implemented in Simio as nodes, however the type of node selected was different and more basic compared to that used for incident locations mainly because no transport logic (i.e. picking up of patient entities) was required. Holding point nodes were placed in areas identified in Table 3.9, approximated in the facility window grid described above. In some cases, specific holding point locations were specified (e.g. at a community health clinic) and nodes were placed at corresponding locations in the facility window. In other cases, where no specific location could be determined, holding point nodes were placed approximately at the centroid of the relevant area or suburb, by visually approximating this location in the facility window.

Allocation of Ambulances to Holding Points

At the start of each simulation run, ambulances were allocated to each holding point in accordance with allocation data in Table 3.9. A reference to each holding point was placed in a data table and each ambulance in a sector group was transferred to the relevant holding point at run initialisation (i.e. the "set-up" period before the simulation run begins and any data are generated).

Movement of Ambulances Between Holding Points

In order to facilitate demand-based movement of ambulances between holding points, each holding point was associated with a state variable to track its capacity (i.e. the number of available

ambulances at that holding point). This state variable was updated every time an ambulance was allocated or has its capacity released. Monitors were then associated with each state variable and set to execute a process each time the state variable's value decreased to zero, or increased above one.

The processes linked to monitors described above contained logic to search other holding points in the sector, determine which was the closest with more than one unit of capacity (i.e. one available vehicle) and move an ambulance to a holding point with zero capacity. When that holding point's capacity was restored by its ambulances becoming available again, a corresponding process was executed (by a monitor) that caused the "borrowed" ambulance to move back to its original holding point. In the event that no holding point could be identified with available capacity, no movement of ambulances took place.

3.6.2.6. Hospitals

Hospitals as Nodes

Hospitals were also implemented as nodes, with similar capabilities as incident nodes. Process logic was added to the *on_enter* event of these nodes in order to customise behaviour of ESVs when entering them, which simulated arrival at the receiving hospital.

Hospital Waiting Times

The delay occurring at hospitals was modelled in accordance with the system data shown in Table 3.7. Ambulances were held at this delay, following which their patient entities were unloaded and destroyed at the hospital node (see below) using process logic. In cases where a PRV had assisted transport, the PRV was forced to wait with its ambulance at the hospital node until the patient entity had been unloaded and disposed of.

Disposal of Patient Entities

Patient entities unloaded from ambulances at each hospital node were destroyed using process logic as soon as possible. This brought about the release of any resource capacity held by patient entities (ambulances or a PRV), allowing them to be allocated immediately to another waiting patient entity, if one existed.

3.7. VERIFICATION PROCEDURES

Verification methods applicable to simulation models have been discussed in Chapter 2. Several of these were adopted and are described below.

3.7.1. Modular Development and Verification

The approach followed in developing the computer model was to begin with small, isolated units of functionality that could be easily verified, using many of the approaches discussed below (e.g. animation, checking of output data and use of debugging tools). To each of these small "sub-models", more elaborate logic and structure was added, and then tested and verified. This lead to the development of a single sector with verified behaviour of all objects and their relationships, followed by expansion to the modelled system in its entirety, comprised of four sectors as described below.

Initially, a small-scale model with 20 incident nodes, two hospitals and three transporters (i.e. ambulances) was used to ensure that the basic behaviour depicted in Figs 3.2 - 3.5 with regard to patient entity transfer to incident nodes, reservation of ambulances, loading of patient entities, routing to a hospital and unloading of patient entities was correct. Approximate arrival rates and distributions were used, and no attempt was made at this stage to create realistic spatial relationships between incident nodes and hospitals. Once behaviour of the isolated unit of functionality above had been verified, PRVs were added along with modelling of additional related behaviour.

The limited test model described above was then expanded to include additional incident nodes, to the scale of Sector 1 in the existing model (one of two smaller sectors), and only one hospital. Holding point nodes were added and non-response movement of ambulances between holding points was implemented and verified for correct behaviour. At this stage incident nodes were placed in approximate positions in the facility window and patient entities were routed to them randomly without attention to realistic proportional distribution.

In the last phase of model development, incident node modelling was carried out as described in 3.6.2.1, across all four sectors. Once completed, a single sector (Sector 1) with a full complement of ESVs was used to generate response time, ESV availability and other output data. Patient entity arrival rates were varied using a scaling factor setting in the patient entity source configuration, and the effects of both increasing and decreasing arrival rates on response times and ESV availability was

assessed and verified against predicted effects of these changes. Once this was completed, all four sectors were fully implemented by replication of process logic and all other objects and settings as required and in accordance with input data. Assessment of output data across a range of patient arrival rates was repeated.

The above process of unitary development and incremental scaling of model complexity and features proved to be an effective and manageable way of conducting model verification, detecting inconsistencies early and correcting the required aspects of the computer model's object configurations and settings, spatial relationships and process logic.

3.7.2. Use of Simio's Graphical User Interface and Development Tools

Model development in Simio does not involve the writing of any programming code. Rather, model objects and elements are presented by means of a rich GUI with extensive configuration options. The facility window provides a two-dimensional space in which model objects such as vehicles and nodes can be configured and positioned as required. Process logic is constructed using a flow chart-like process development window. The visual nature of Simio's development environment and avoidance of programming code as a way of constructing models lends itself to easier understanding of the model and its processes.

3.7.3. Animation and Observation as a Verification Tool

Simio uses animation as an integral part of the model development environment, meaning that no additional effort is required in order to create animated views of a model while it is running, in two- or three-dimensions. The use of labels attached to animated objects was particularly helpful in verification, as any of an object's states or properties could be displayed in the label and observed as the animation progressed.

Animation was used extensively to check model behaviour at all stages of development, particularly with regard to the verification of specific behaviour represented in Figs 3.2 - 3.8, and after the integration of smaller parts of the model into a larger-scale one (as described in 3.7.1). Incorrect or suspicious behaviour of objects (ESVs in particular) could be identified, confirmed and observed any number of times at any speed. This was often the first step in either correcting an assumption or logic in the conceptual model, or changing part of the computer model. An example of animation in Simio's user interface is given in Annexure N.

3.7.4. Checking of Output Data

Collection and analysis of output data in Simio was easy to accomplish. Once a tally statistic for a specific variable had been defined, data (counts, if applicable, average, maximum and minimum) were automatically reported in a pivot grid at the end of each simulation run. Writing of output data to files aided in more detailed analysis if required.

Checking of output data was performed as a method of verification throughout development of the model, however more weight was attached to this as development progressed and its use was quite limited in the very early stages of simple, unitary model development. Response time outputs were assessed from the point of having all four sectors implemented as prior to that spatial relationships between incident nodes, and between incident nodes and hospital nodes, had not been finally determined. Response times and ESV availability were assessed for changes in response to increased patient arrival frequency, to establish whether changes that occurred were reasonable.

Time intervals derived from distributions (such as dispatch times and scene times) were collected in an output text file and then analysed to determine whether their distribution did in fact match the specified distribution in the model. Similarly, counts of patient entity arrivals at each incident node were written to an output text file. These data were compared to proportional allocation of patient entities as determined in the model specification, to ensure that there was agreement between "busy" and "quiet" areas in the real system and the model.

3.7.5. Use of Simio's Debugging Tools: The Model Trace and Watch Facility

Simio provides a sophisticated set of debugging tools which make the identification and correction of errors and anomalous model behaviour relatively easy. Two will be discussed in the context of verification, as they were of great value in this process in every stage of model development.

3.7.5.1. The Model Trace

Simio's trace can be enhanced with the addition of breakpoints – points in the model's process logic where execution can be stopped (and restarted if desired). The level of detail provided in Simio's trace aided greatly in being able to track model execution in fine detail, event by event, and also to diagnose the source of errors when they occurred. Trace output was used extensively throughout model development to ensure that the intended model behaviour was accompanied by expected events and responses, state changes, and process execution in a variety of conditions.

3.7.5.2. The Watch Facility

In Simio, the Watch facility is a window containing a listing of all model objects that are being "watched" - meaning that they have been included in the set of objects whose properties and states can be viewed whenever model execution is paused during a simulation run. This is different when compared to the model trace above, as the trace sequentially records only the active states of a sequence of executing processes while any property or state of a watched object can be viewed when model execution is paused. In addition to objects, global elements of the model can also be viewed as part of the watch window.

Like the trace facility however, the Watch facility is an extremely powerful tool as it provides a detailed snapshot of any object at any time during the execution of a simulation run. This aids in the debugging process, where the source of an error can often be located amongst the state variables of one or more objects. But even more importantly, correct behaviour of any object can be verified by repeatedly viewing object properties, states, queues and other data sequentially as the object proceeds through the simulation and interacts with other objects.

3.8. VALIDATION PROCEDURES

Model validation was carried out in line with the recommendations given in Chapter 2, and a combination of white- and black-box validation were performed. White-box validation was performed throughout the development process, as components of the model were completed. Functionality of components was checked using a variety of methods described above including output reports and animation.

In several places throughout the model values for data selection were specified. For example, values for proportional allocation of patient entity state variables, such as those indicating patient priority or exempt status. Allocation of state variables was implemented by placing the required state names and proportions in a data table and allowing the simulation software to assign each state name to new patient entities in the desired proportions (as specified in the table).

In order to ensure that the model was producing valid data, in line with the chosen settings, patient entity state variables were written to output files during simulation execution. These data were compared to the required settings at various times in the model development process. Comparisons of output data and required values or distributions are shown in 3.8.1 – 3.8.2 below, calculated over 15 replications.

3.8.1. Proportional Distribution: Patient and Incident States

Patient entity states are shown in Table 3.12. Required values, that is values for each patient entity state variable set in the simulation's configuration, and values actually produced by the simulation after 15 replications are given for the purposes of comparison.

Table 3.12. Patient and Incident States: Required and Actual Values

Sector	Patient Entity State	Required (Set) Value	Output Value		Mean Difference
			Mean	95% CI	
1	Priority	P1 = 50%	49.66%	49.58; 49.74	-0.3
		P2 = 50%	50.34%	50.26; 50.42	0.3
	Exempt	Exempt P1 = 30%	30.18%	30.11; 30.26	0.2
		Exempt P2 = 30%	30.56%	30.48; 30.63	0.6
2	Priority	P1 = 45%	44.48%	44.43; 44.54	-0.5
		P2 = 55%	55.52%	55.47; 55.56	0.5
	Exempt	Exempt P1 = 30%	30.59%	30.54; 30.64	0.6
		Exempt P2 = 32%	31.23%	31.18; 31.27	-0.8
3	Priority	P1 = 43%	42.20%	42.16; 42.25	-0.8
		P2 = 57%	57.80%	57.76; 57.83	0.8
	Exempt	Exempt P1 = 34%	34.45%	34.41; 34.49	0.4
		Exempt P2 = 38%	38.73%	38.70; 38.77	0.7
4	Priority	P1 = 45%	43.81%	43.74; 43.87	-1.2
		P2 = 55%	56.19%	56.13; 56.25	1.2
	Exempt	Exempt P1 = 35%	33.87%	33.80; 33.93	-1.1
		Exempt P2 = 36%	35.7%	35.64; 35.76	-0.3

95% CI = 95% Confidence Interval, P1 = Priority 1, P2 = Priority 2

The differences between required proportions and those generated by the simulation were small.

3.8.2. Proportional Incident Node Patient Allocation

The method used to distribute patient entities amongst incident nodes in the simulation model was described in 3.6.2.1. Each incident node was entered into a data table (separate tables were used per sector) and in the same table row a number was added indicating that node's proportional incident weighting. As patient entities were created during a simulation run, process logic was used to randomly allocate each new entity to a table row (i.e. an incident location). This was done by

using the *RandomRow* function in Simio, which also takes into account the row's proportional incident weighting and selects rows in a similar proportional distribution. As part of the validation process, each patient entity's incident node allocation was written to a file during the simulation run. These data are compared to the required proportional allocations using mean differences and correlation coefficients as shown below in Table 3.13.

Table 3.13. Incident Node Allocations: Required and Actual Values

Sector	Mean Difference*	Correlation Coefficient	p
1	0.146	0.996	< 0.001
2	0.091	0.999	< 0.001
3	0.044	0.998	< 0.001
4	0.078	0.995	< 0.001

*Between required and actual proportional allocations

Small mean differences between required and actual incident node allocation in all sectors, together with very strong correlation indicate good agreement between the spatial distribution of incidents as observed in the real system and as modelled. This was also confirmed in a less quantitative way during verification and validation, by observing the relative frequency with which vehicles visited nodes to pick up patients in each sector.

3.8.3. Time Interval Comparisons

The final quantitative form of validation performed was black-box validation of the model once it was completed, comparing response time and other time interval outputs with real system data. This was done in two ways; by plotting means and 95% confidence intervals of system and simulation times and comparing them visually, and by constructing 95% confidence intervals for the mean difference between system and simulation times as described by Law and Kelton.(81)

Values for the system were derived from random samples taken from the 2012 set of response data. These samples ware separate from the sample used for input analysis, which was also from the 2012 data set. The decision to use data from the same year as that used for input analysis was based on the observation that there were differences in all of the time intervals used for validation between the 2012 and 2011 data sets ranging from 0.57 to 6.57 minutes.

Eight charts (Figs 3.9 – 3.16) show the comparative mean system and simulation P1 and P2 response times, scene times and transport times. These are followed by confidence interval comparisons of the same time intervals in Table 3.14.

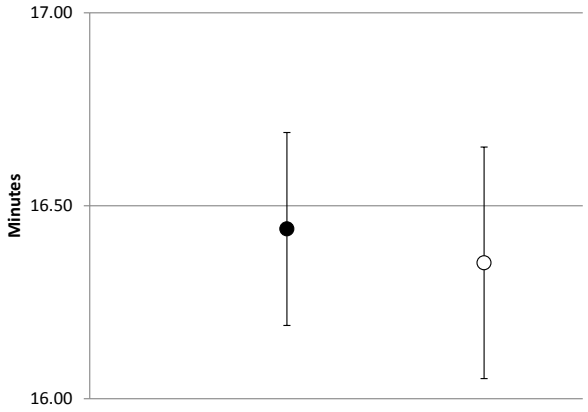


Figure 3.9. Priority 1 Total Response Time

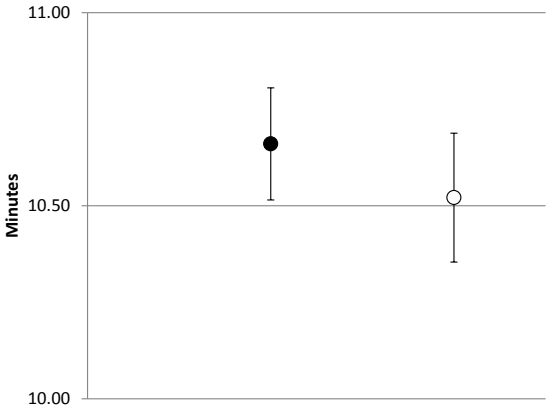


Figure 3.10. Priority 1 Travel Response Time

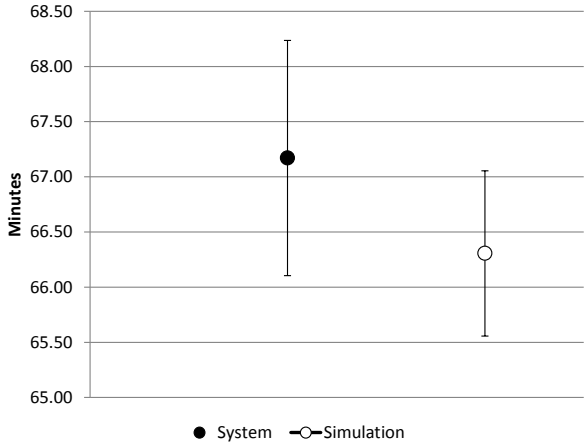


Figure 3.11. Priority 2 Total Response Time

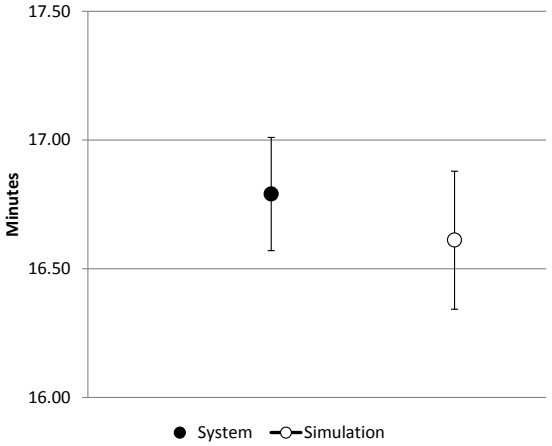


Figure 3.12. Priority 2 Travel Response Time

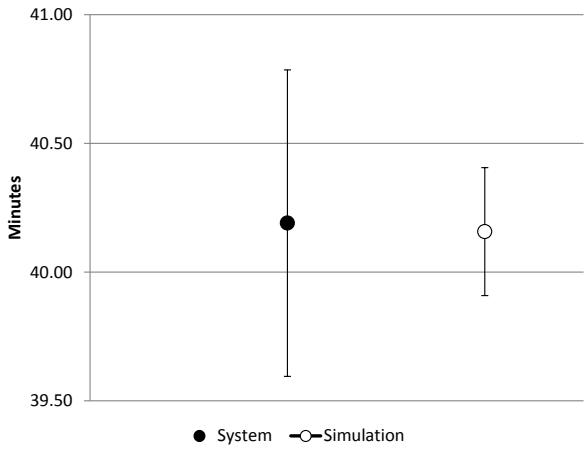


Figure 3.13. Priority 1 Scene Time

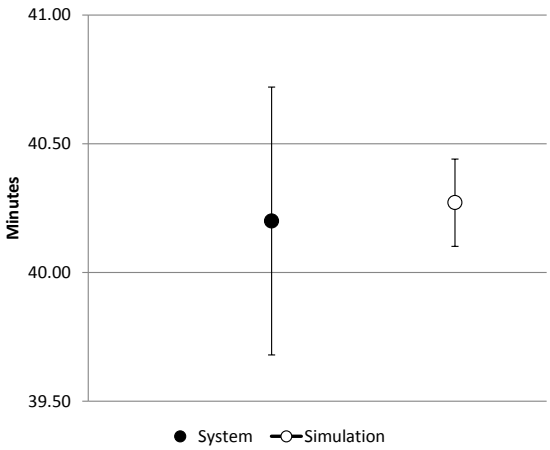


Figure 3.14. Priority 2 Scene Time

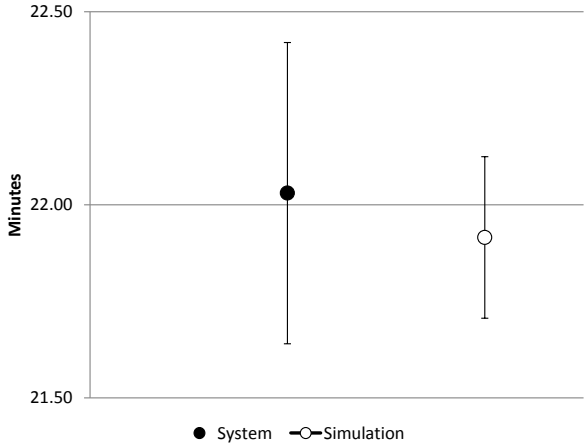


Figure 3.15. Priority 1 Transport Time

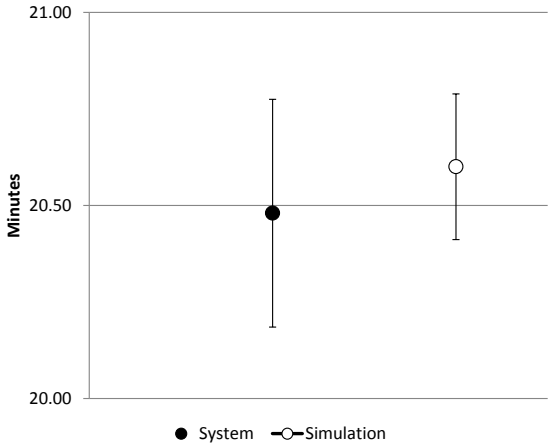


Figure 3.16. Priority 2 Transport Time

Visual assessment of the charted mean values and confidence intervals above confirms little difference between system and simulation outputs. P1 response times (total and travel) and both scene times (P1 and P2) were more closely aligned than P2 response times and P1 and P2 transport times. Dispersion in system data was generally greater than in simulated data (as evidenced by wider confidence intervals) except in the P1 response time groups.

The confidence intervals shown in Table 3.14 were constructed using the technique described by Welch as set out in Law and Kelton.(81) This technique is suitable for the comparison of system and simulation outputs where sample sizes of the two groups are unequal.

Table 3.14. Response Intervals: System vs. Simulation

Time Interval	Mean Difference	% Difference	95% CI for Difference
P1 Total Response Time	0.088	0.536	-0.061; 0.237
P1 Travel Response Time	0.139	1.313	0.101; 0.177*
P2 Total Response Time	0.864	1.295	0.635; 1.093*
P2 Travel Response Time	0.179	1.072	0.119; 0.239*
P1 Scene Time	0.033	0.082	-0.076; 0.142
P2 Scene Time	-0.071	0.177	-0.160; 0.018
P1 Transport Time	0.115	0.522	0.040; 0.190*
P2 Transport Time	-0.120	0.585	-0.179; -0.061

95% CI = 95% Confidence Interval, P1 = Priority 1, P2 = Priority 2, * = Significant Difference

As with all confidence intervals for a difference between means, interpretation of a difference between the groups (in this case system and simulation) involves consideration of whether the relevant confidence interval contains zero. If it does, then it cannot be ruled out that the mean difference between the groups is zero and thus the groups are considered to be equivalent. From a validation perspective this is desirable as it suggests that system and simulation outputs are the same.

Confidence intervals for P1 total response times contain zero as do those for both scene times. These three groups can thus be considered equivalent or not significantly different. The remaining groups have confidence intervals excluding zero and thus statistically are significantly different. In such cases, Law and Kelton recommend assessing the percentage difference in order to form an opinion regarding the practical significance of the difference between means. If it is unlikely that the percentage difference would be of any practical significance then it may be acceptable to say that the mean difference is small enough to judge system and simulation outputs to be equivalent.(8)

Although the two P2 response time groups, P1 travel response time and the two transport time groups have mean times that are by definition significantly different, the percentage differences are very small ranging between 0.082 and 1.313. These differences are unlikely to be of practical significance in drawing conclusions from the model. Consequently, both visual inspection of the time interval comparisons between system and simulation above and consideration of confidence intervals for the difference between means, support the assertion that the simulation model is a close enough approximation to the system for the purposes of this study.

3.8.4. Turing Test

As an additional step in validation of the model, the opinions of two system experts were sought in the form of a Turing test, as described by Law and Kelton and Robinson.(8,75) A Turing test involves presenting system experts with two sets of summarised quantitative outputs; one set derived from the real system and one set derived from the simulation. These data are presented in such a way that, apart from the numerical values, they are indistinguishable (e.g. the two reports should be formatted and presented in an identical way). The system experts are asked to consider both sets of data and to attempt to distinguish between them and identify the source of each. If this is not possible, it is seen as enhancing credibility of the model. If it is possible for the system experts to correctly distinguish between the data sets, the basis of this can be used to improve the model or better understand it.

Two system experts from the Western Cape Emergency Medical Services with extensive knowledge and experience of the dispatch system and EMS operations were selected. Both were presented (individually and on separate occasions) with two identically formatted reports containing response and other time interval summaries for P1 and P2 responses (as shown in Annexure E) and asked to distinguish between them as outlined above. The system data for these reports were taken from the set of data used to validate the model, as described in 3.8.3.

Both experts were able to correctly differentiate between system and simulation data sets on the basis of a single statistic, namely the proportion of P2 responses within 60 minutes (63% for the system data and 84% for the simulation data). When asked how the difference in these statistics allowed them to correctly identify the system data set, their responses both centred on the efficiency of dispatchers in dealing with P2 cases.

Both experts described how, according to their knowledge and experience with the dispatch system, dispatchers tend to focus most of their attention on time frames with P1 cases because of the emphasis placed on this in quality benchmarking. It was the opinion of both experts that, given the large number of P2 responses and the additional responsibility of co-ordinating resources for inter-hospital transfers (which were not considered as part of the response system in this study), the efficiency of most dispatchers will tend to be better for P1 cases than for P2. Both experts agreed that the 21% higher within-target statistic for P2 responses in the model output could be explained by the fact that efficiency fluctuations due to human factors will not be seen in a simulation where the efficiency of algorithmic, rule-based execution of processes by a computer is constant.

3.8.5. Experimentation Validation

The three questions related to experimentation validation discussed in 2.5.3.4 are discussed in this Section. These questions relate to the need for and determination of a warm-up period for the simulation, and choice of the simulation run length and number of replications.

3.8.5.1. Warm-up Period

In order to decide whether a warm-up period was required, the approach recommended by Robinson was used.(80) This utilises a time series plot as initially described by Welch in order to identify model output which may be subject to initialisation bias, and to gain an impression of when this period of initialisation bias ends in order to identify the warm-up period. Using this approach,

the smallest window size for moving averages is selected that produces a smooth line. Combined P1 and P2 travel and total response time data generated by selecting 15 replications of a 60-day run length are shown in Fig 3.17 and Fig 3.18, using a window size of six (30 of the 60 days are shown).

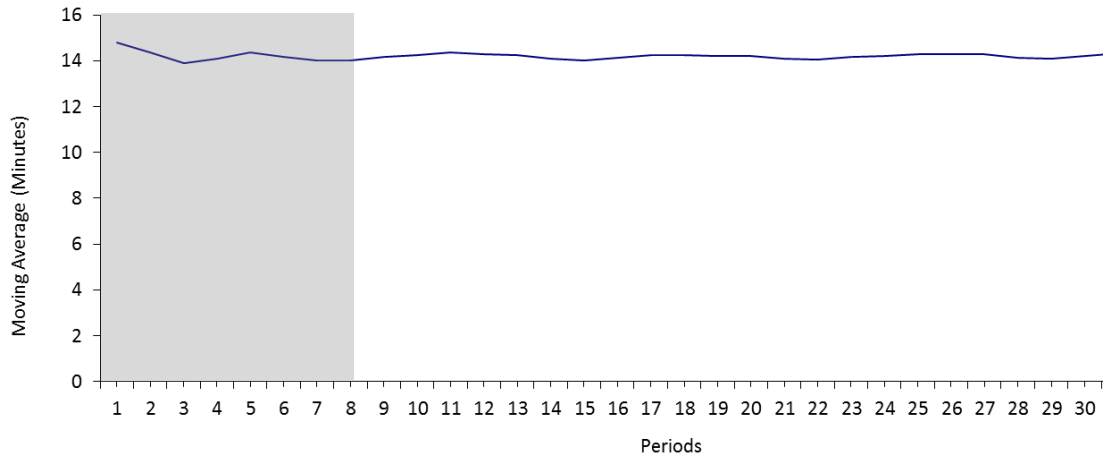


Figure 3.17. Time Series: Aggregated P1 & P2 Travel Response Time

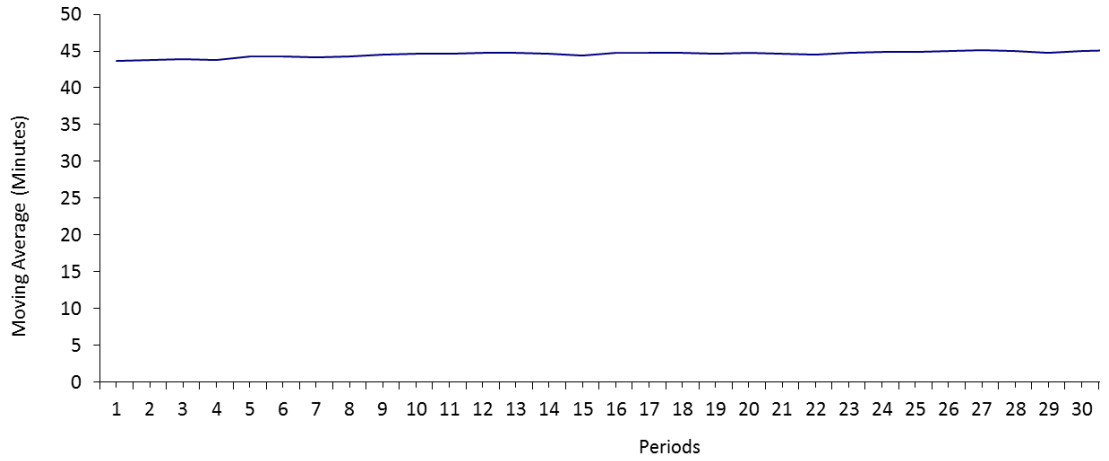


Figure 3.18. Time Series: Aggregated P1 & P2 Total Response Time

Some evidence of initialisation bias is seen in Fig 3.17, within the first seven to eight days while Fig 3.18 shows no sign of any initialisation bias with the selected window size. Consequently, the warm-up period was chosen as eight days (the shaded area in Fig 3.17).

3.8.5.2. Run-length and Replications

Changing the time series plots above to have a window length of five shows a clear seven-day steady state cycle in both travel and total response time data (Figs 3.19 and 3.20 below).

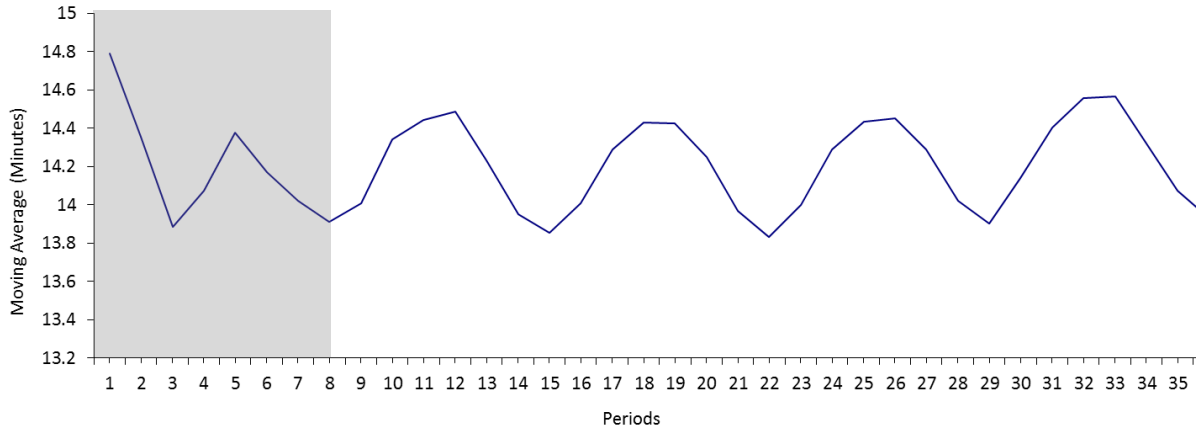


Figure 3.19. Time Series: Aggregated P1 & P2 Travel Response Time (window = 5)

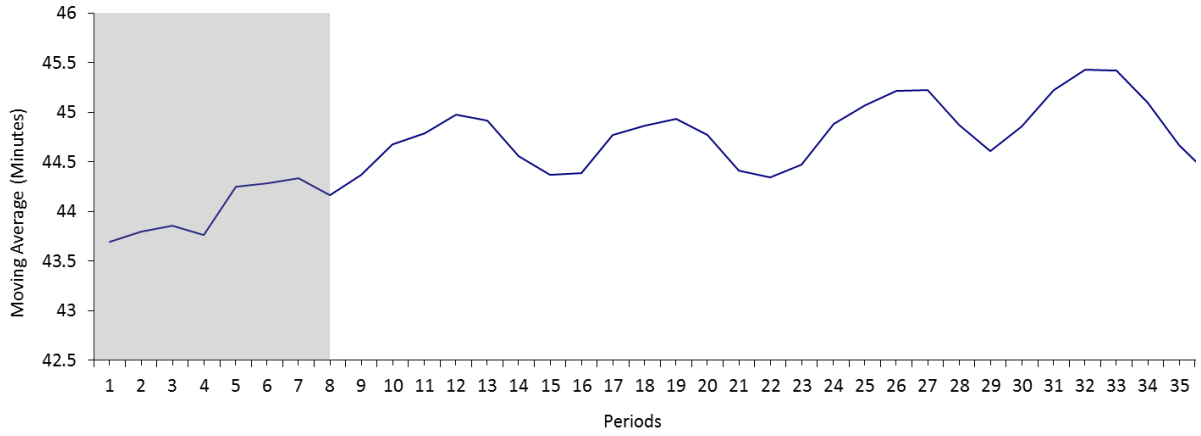


Figure 3.20. Time Series: Aggregated P1 & P2 Total Response Time (window = 5)

Taking this steady state cycle into account, a simulation run length of seven days was chosen. In order to assess the required number of replications, plots of the cumulative means and 95% confidence intervals by replication number were constructed for aggregated P1 and P2 travel (Fig 3.21) and total response time data (Fig 3.22), as described by Robinson.(80) Percentage deviation from the confidence interval diminishes until a final value after 15 replications of 0.43% for travel response time and 0.45% for total response time. This was considered adequate for the study objectives and consequently a setting of 15 replications was chosen for experimentation.

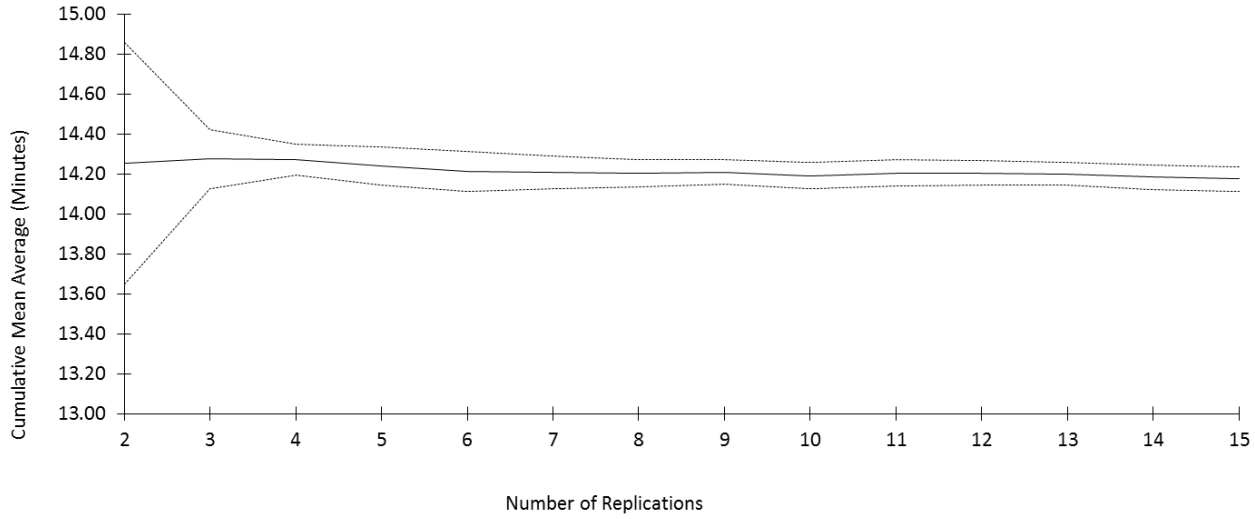


Figure 3.21. Cumulative Mean & 95% Confidence Interval: P1 & P2 Travel Response Time

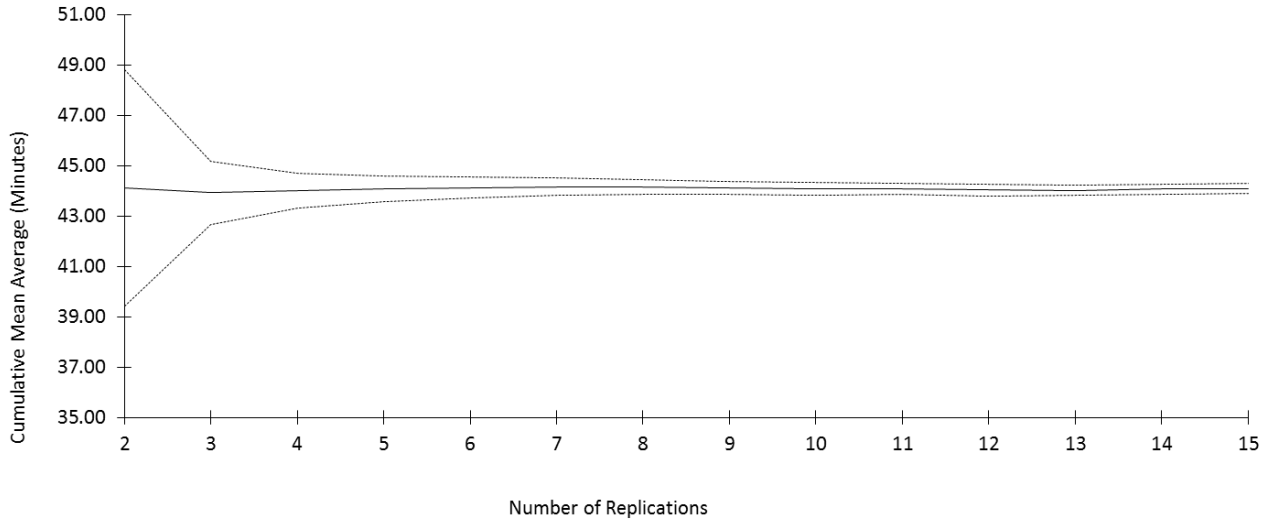


Figure 3.22. Cumulative Mean & 95% Confidence Interval: P1 & P2 Total Response Time

3.9. CHANGES TO THE VALIDATED MODEL: IMPLEMENTATION OF THE EXPERIMENTAL FACTORS

As initially described in the study objectives, and in 3.4.2.6 in this Chapter, the validated simulation model was changed in order to implement two experimental factors. These changes are described below.

3.9.1. Response Model: Single- and Two-Tier

In order to implement single- and two-tier response models, the behaviour of PRVs in the validated model was modified as described in 3.4.2.6. For a single-tier response model, PRVs were retained

but were disabled through changes in process logic. To keep the balance of ALS and non-ALS resources similar to that in the two-tier response model, an additional ALS ambulance was added in each sector to replace the disabled PRVs. For the two-tier response model, the opposite approach was taken and process logic was modified to facilitate the dispatch of a PRV to every P1 incident, with a set of restrictions as described in 3.4.2.6.

3.9.2. Vehicle Location: Static and Dynamic

Implementation of the dynamic vehicle location model required no changes as this approach was identical to the one used in the validated model. For the static vehicle location model holding points were retained but were disabled through process logic and consequently any movement of vehicles between holding points as described in 3.6.2.5 was also disabled. At the initialisation of every replication vehicles were located at the hospital in their sector and returned to this location when idle during the simulation run. Vehicle numbers, and relative ALS and non-ALS vehicle numbers were not changed beyond the requirements for the two different response model configurations as described in 3.9.1.

3.9.3. Selection of Locations for Additional Vehicles

The last objective of this study, as set out in 1.6.3 of Chapter 1, was to assess the effect of increased vehicle numbers on response performance of the best-performing simulation model (assuming that response target benchmarks had not already been met by this model). In order to do this, vehicle numbers had to be increased in an incremental way with each new model scenario, across all four sectors.

Logically, the effect of increasing vehicle numbers as identified above could be influenced by where exactly those vehicles are placed. After all, the location of vehicles was selected as an experimental factor in this study because it was thought to influence response performance. In order to implement the increasing of vehicle numbers in a way that augments vehicle location logic already present in each of the models, the following approach was used.

3.9.3.1. Static Vehicle Location

In the case of models with static vehicle location, additional vehicles were added to each vehicle group located at the hospital in each sector.

3.9.3.2. Dynamic Vehicle Location

Models with dynamic vehicle location presented more of a problem than those statically located, as areas of demand relative to holding points could alter response performance depending on where additional vehicles were placed. In order to maximise the effect of additional vehicles in a consistent way, vehicle placement was implemented as follows:

- A state variable associated with each holding point was created in the baseline model. Each of these state variables tracked the capacity (i.e. number of available vehicles) at its allocated holding point. A monitor object was attached to each of these state variables and was set to trigger a process whenever capacity of the corresponding holding point dropped to zero, recording the system time. A separate monitor triggered another process when the opposite occurred (the holding point's capacity rose above zero) and this process tallied the difference between the current system time and the previously recorded beginning of zero capacity time. In this way, each interval of zero capacity during a simulation run was tallied for each holding point, and the average of these was available after the run as an output statistic.
- A data table was created in each of the four experimental models containing a row for each holding point (in each sector) and a column containing a weighting derived from the zero capacity data mentioned above. This weighting represented the proportional zero capacity time for each holding point per sector and was thus an approximation of incident distribution and response frequency in the area around the holding point.
- During experimentation, when vehicle numbers were increased, process logic was used to search the tables referred to above (per sector) and proportionally allocate vehicles over-and-above those in the baseline setup to holding points based on the weightings described.

The effect of this approach was that as vehicle numbers were increased above baseline, these added vehicles were predominantly located at holding points with the greatest historical zero capacity time, where they would be expected to have the most significant effect on response performance by increasing capacity. Most importantly, this demand-based location of additional vehicles was applied in the same way to each of the different models in order to eliminate variance in this aspect of model behaviour.

3.10. DATA ANALYSIS

Data analysis focused on two main objectives, as set out in 1.6.2 (iii) and 1.6.3 of Chapter 1. The first objective was to compare total response times from the four different models described above,

representing all possible combinations of the two experimental factors, and to identify a "best" model based on response time performance. The second objective was to assess the effect of increased vehicle numbers on the response time performance of the "best" model, and to optimise this performance. Data analysis for the first objective is described in 3.10.1 below, while that for the second objective is described in 3.10.2.

3.10.1. Experimental Design and Hypotheses

Following the terminology used by Law and Kelton,(81) a 2^k factorial design was used. With two experimental factors each having two levels, this required four different scenarios for each dependent variable set, as indicated in Table 3.15.

Table 3.15. Experimental Factors and Responses

Scenario (Model)	Response Model	Vehicle Location	Responses*
1	Single-tier	Dynamic	P1 Response Time; P2 Response Time; Proportion of P1 and P2 Responses Within Target, Vehicle Availability
2	Two-tier	Dynamic	
3	Single-tier	Static	
4	Two-tier	Static	

* Response times are total response times, P1 = Priority 1, P2 = Priority 2

Implementing process logic and other changes to effect the factorial changes in a single model was not considered as this would have increased complexity of the model significantly. Rather, four different models were used, with changes from the validated model as described in 3.9. Statistical hypotheses for dependent variable means were H_0 that $\mu_1 = \mu_2 = \mu_3 = \mu_4$ and H_A that population means were not equivalent.

3.10.2. Changes to Vehicle Numbers and Optimisation

The effect of increases in vehicle numbers was evaluated in two different ways. The first was a non-optimised approach where vehicle numbers were doubled from the baseline values for the best model with successive scenarios and changes in the proportions of P1 and P2 responses meeting target values were observed. These increases were continued until either the response target values were met or no further change was observed.

Subsequent to this, optimisation software was used in order to obtain a more fine-grained impression of the best combination of vehicles (across the four sectors, and ALS vs. non-ALS

vehicles) to maximise the proportions of P1 and P2 responses meeting target values within the upper limit obtained from the non-optimised approach.

3.10.3. Statistical Procedures and Software

Multivariate analysis of variance was used for statistical analysis of output from the experimental design described above. The General Linear Model (multivariate) procedure in IBM SPSS (version 22, IBM Corporation, New York, USA) was used with a 5% significance level. Contrasts were selected in order to compare different levels of each factor and interaction plots were produced of estimated marginal means for all dependent variables across both factor groups. In order to assess fit of the linear model with experimental data residuals were checked for mean values, constancy of variance and normality. Residuals were also checked for signs of autocorrelation.

Optimisation was carried out with OptQuest software (version 6.6, OptTek Systems Inc. Colorado, USA) which was available as an add-in from the Simio Experiment Tools window. The optimisation objective type was set to multi-objective weighted for the two outputs of P1 responses within 15 minutes and P2 responses within 60 minutes. Both of these were formatted as proportions with lower bounds of 0.9 and 1.0 respectively and objective set to maximise. Optimisation weightings of 0.75 and 0.25 were applied respectively, reflecting the fact that meeting P1 response targets is considered to be of greater importance than meeting P2 response targets for optimisation purposes.

3.11. SUMMARY

In this Chapter a detailed description of the chosen modelling approach was presented, followed by an explanation of the data analysis method . In describing the baseline conceptual model, the framework discussed in Chapter 2 was used to set out the modelling objectives, inputs and outputs, model content, process logic and assumptions and simplifications. Additional process logic, assumptions and simplifications for the extended models (representing combinations of the experimental factors) were also explained.

Description of input data sources and analysis provided a basis for understanding how key aspects such as incident arrival rates, spatial distributions and ESV location and numbers were modelled. This set the scene for a detailed account of exactly how the conceptual model was translated into a software representation, both for the baseline model and for the extended experimental factor models. Credibility of the computer model was addressed through a thorough description of verification and validation procedures, many of which were presented in Chapter 2 as approached

recommended in the literature. Many of the chosen simulation software application's features were described, both in the account of how model translation was accomplished and in the account of verification and validation procedures.

The description of data analysis methods given in the latter part of this chapter emphasised the experimental nature of the study, and the approach to quantifying the effects of each of the identified experimental factors. The method employed to determine the effect of ESV numbers on response performance was explained by referring to the optimisation software used and how this was configured.

CHAPTER 4: RESULTS

4.1. INTRODUCTION

The four computer simulation models used in this study, described in detail in the previous chapter, each represented a combination of two levels of two experimental factors - ESV location and response model. Each of these models was run for a period of seven days of simulated time, for 15 replications to produce output data for analysis. The primary dependent variables were P1 and P2 response time and the proportion of P1 and P1 responses meeting previously defined response time targets.

In this Chapter, descriptive results for the four model outputs are first presented showing mean response time and proportional response time target compliance values for each factor level combination. This is followed by results of hypothesis tests assessing the effects of ESV location and response model on response time and proportional response time target compliance. Using the results of these tests, a best-performing model is identified.

The impact of ESV numbers on the best-performing model is assessed by adding additional ESVs first without optimisation and secondly using optimisation. The effect of ESV location and response model on ESV availability is described, together with values for the hand-off delay across ESV location and response model factors. Finally, additional data related to the primary response time and proportional response time target compliance variables are described including data on the availability of ALS at incidents, response distances, cross-sector responses and mission times.

4.2. THE EMERGENCY RESPONSE INTERVAL: COMPONENT VALUES FOR ALL FACTOR MODELS

Mean values for each component of the ERI, as depicted in Fig 1.1, are shown in Figs 4.1 and 4.2 (on the next two pages), separated by factor model. Different charts are presented for P1 and P2 responses.

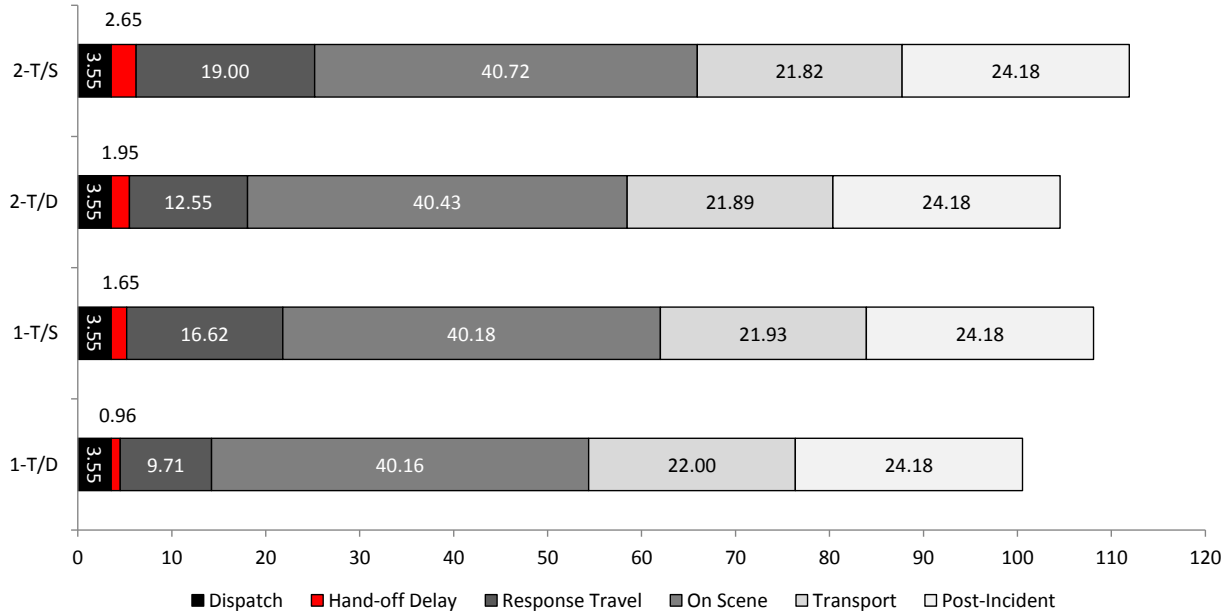


Figure 4.1. Components of the Emergency Response Interval: Priority 1 Responses
All values are means in minutes, 1-T/D = single-tier dynamic, 1-T/S = single-tier static, 2-T/D = two-tier dynamic, 2-T/S = two-tier static

Mean values for the dispatch and post-incident intervals in Figs 4.1 and 4.2 are the same, as these intervals were derived from the same single probability distribution across all factor models and the use of common random numbers removed variation in these intervals between the four models. Although a similar approach was used to determine the on-scene interval, variation between these intervals across factor models stems from the way in which on-scene delays for exempt incidents were modelled, as these were not determined by a single probability distribution. Values for the remaining intervals were determined by travel times and were subject to variation as an effect of the interplay between ESV location and the type of ESV allocated, as determined by the combination of experimental factors in each case.

In Fig 4.1 the hand-off delay is the mean time interval between completion of the call taking and dispatch process (i.e. when a dispatcher is ready to “hand off” an incident to an available ESV) and when an available ESV receives the incident. The implications of this delay, and factors affecting it, are set out in more detail in Section 4.5 of this Chapter.

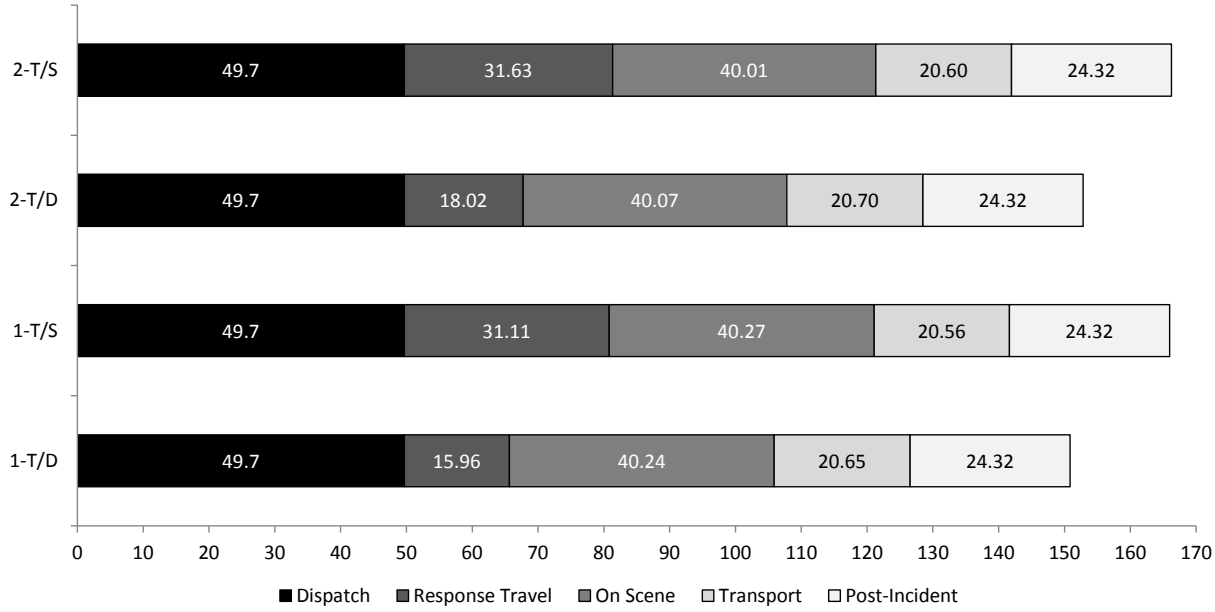


Figure 4.2. Components of the Emergency Response Interval: Priority 2 Responses
All values are means in minutes, 1-T/D = single-tier dynamic, 1-T/S = single-tier static, 2-T/D = two-tier dynamic, 2-T/S = two-tier static

In both of Figs 4.1 and 4.2 the effect of each factor model on response travel time, and by extension on mission time, can be seen. These effects are analysed probabilistically in 4.3.2. Hand-off delays for P2 responses were very small and are thus omitted from Fig 4.2 (see Section 4.5 of this Chapter for the range of P2 hand-off delay values).

4.3. RESEARCH OBJECTIVE 2 (iii): DIFFERENCES BETWEEN EXPERIMENTAL FACTORS: RESPONSE MODEL AND VEHICLE LOCATION

Data on total response times (referred to below as 'response times') and proportions of responses meeting total response time targets were generated by running each model representing one of the four factor level combinations for a simulated time of seven days, over 15 replications. Data from each replication were written out to separate text files as the replication was executed and then transferred to a single spread sheet file. These aggregated data were imported into IBM SPSS for analysis.

4.3.1. Descriptive Data

Descriptive data for P1 and P2 response times across each factor level combination are shown in Table 4.1. Each mean is itself composed of the mean response times for 15 replications.

Table 4.1. Descriptive Data: P1 and P2 Response Times

Response Model	Vehicle Location	Group	Mean (min.)	95% CI
Single-tier	Dynamic	P1	14.22	13.95; 14.50
		P2	65.66	64.92; 66.40
Single-tier	Static	P1	21.82	21.59; 22.04
		P2	80.84	80.16; 81.52
Two-tier	Dynamic	P1	18.05	17.61; 18.49
		P2	67.72	67.01; 68.44
Two-tier	Static	P1	25.20	24.85; 25.54
		P2	81.33	80.56; 82.09

95% CI = 95% confidence interval, P1 = Priority 1, P2 = Priority 2

Static vehicle positioning results in longer response times for both P1 and P2 responses, across both levels of response model. A single-tier response model with dynamic vehicle positioning yielded the best overall P1 and P2 response performance, with a 24% shorter P1 response time and a 3% shorter P2 response time compared to a two-tier response with dynamic vehicle positioning.

Descriptive data for proportions of P1 and P2 responses meeting their respective response targets are shown in Table 4.2 below. Targets were P1 responses within 15 minutes and P2 responses within 60 minutes.

Table 4.2. Descriptive Data: P1 and P2 Responses Meeting Response Targets

Response Model	Vehicle Location	Group	Mean (%)	95% CI
Single-tier	Dynamic	P1	68.9	68.18; 69.68
		P2	84.8	84.36; 85.16
Single-tier	Static	P1	37.9	37.48; 38.41
		P2	79.8	79.33; 80.28
Two-tier	Dynamic	P1	63.6	63.05; 64.18
		P2	84.1	83.66; 84.50
Two-tier	Static	P1	37.5	36.75; 38.27
		P2	79.6	79.10; 80.12

95% CI = 95% confidence interval, P1 = Priority 1, P2 = Priority 2

Similar trends to those evident in Table 4.1 can be seen, with dynamic vehicle positioning and single-tier response models accounting of the greatest proportion of P1 and P2 response targets being met.

The effect of this factor combination is most pronounced for P1 responses, with a single-tier dynamic model again yielding the best overall performance. Differences across levels of the response model factor are negligible for both models involving static vehicle locations.

None of the models, representing all combinations of the two experimental factors in this study, managed to achieve the national benchmark of 90% of P1 responses in ≤ 15 minutes and all P2 responses in ≤ 60 minutes. The best performing model fell short of the P1 benchmark by a 27% margin, and the P2 benchmark by a 17% margin. The descriptive data above provide only a crude idea of the differences between experimental groups. Results of inferential statistical analysis on these data are presented below.

4.3.2. Hypothesis Tests

4.3.2.1. Main and Individual Response Effects

Assumptions of multivariate analysis of variance include normal distribution of all dependent variables and homogeneity of variances.(119) The normality assumption was confirmed by assessment of each dependent variable set (Kolmogorov-Smirnov tests, all $p > 0.05$) while the homogeneity of variances assumption was assessed with Box's test of equality of covariance matrices ($M = 30.250$, $p = 0.661$) and Levene's test of equality of error variances (all $p > 0.05$).

Suitability of the statistical model was assessed by analysis of residuals which should display properties of zero mean, constant variance, independence and normal distribution if model fit is considered to be adequate.(119) Residuals for all dependent variables were confirmed to satisfy the first two requirements from descriptive analysis. Partial autocorrelation plots were used to assess the independence requirements. Some evidence was found of autocorrelation in both sets of P2 data (response times and proportion of responses meeting the 60 minute target) suggesting that the model could have been slightly improved. However this is unlikely to have changed any conclusions drawn from the results below. Normality of residuals was assessed with Kolmogorov-Smirnov tests (all $p > 0.05$).Tests of the main effects of the model are shown in Table 4.3 (on the next page).

Table 4.3. Multivariate Test Results

Effect	F	Hypothesis df	Error df	p	Partial η^2
Response Model	156.406	4.00	53.00	<0.001	0.922
Vehicle Location	3493.738	4.00	53.00	<0.001	0.996
Interaction	20.974	4.00	53.00	<0.001	0.613

Interaction = Response Model*Vehicle Location, df = Degrees of Freedom

The multivariate test of overall differences among factors was significant, with moderate to large estimated effect sizes. Observed power (not shown in Table 4.3) was 1.0 for all factors. Values for Hotelling's Trace, which is an estimate of the relative contribution of each factor to the model, were 11.804 (response model), 263.678 (ESV location) and 1.583 (response model * ESV location). Results from univariate between-subjects tests are shown in Table 4.4.

Table 4.4. Univariate Test Results

Source	Dependent Variable	F	df	p	*R^2	Partial η^2
Corrected Model	P1 RT	939.57	3	< 0.001	0.979	0.981
	P2 RT	608.665	3	< 0.001	0.969	0.970
	P1 % Target	3017.540	3	< 0.001	0.994	0.994
	P2 % Target	166.974	3	< 0.001	0.894	0.899
Response Model	P1 RT	543.627	1	< 0.001	-	0.907
	P2 RT	14.158	1	< 0.001	-	0.202
	P1 % Target	90.127	1	< 0.001	-	0.617
	P2 % Target	4.249	1	0.044	-	0.071
Vehicle Location	P1 RT	2272.970	1	< 0.001	-	0.976
	P2 RT	1806.397	1	< 0.001	-	0.970
	P1 % Target	8897.584	1	< 0.001	-	0.994
	P2 % Target	495.391	1	< 0.001	-	0.898
Interaction	P1 RT	2.134	1	0.150	-	0.037
	P2 RT	5.440	1	0.023	-	0.089
	P1 % Target	65.009	1	< 0.001	-	0.537
	P2 % Target	1.282	1	0.262	-	0.022

* Adjusted, RT = Response Time, P1 = Priority 1, P2 = Priority 2, Interaction = Response Model*Vehicle Location, % Target = proportion of cases meeting the relevant response target, df = Degrees of Freedom

The model is significant for all of the dependent variables, as indicated in the each of the dependent variable rows for the corrected model source. Both response model and ESV location have significant effects on all of the dependent variables, except for P1 response time and P2 responses meeting the response target under response model* ESV location. There is some variation in the strength of the effect observed above. Effect sizes for P2 response time and P1 and P2 responses meeting the response targets for the response model factor are smaller than those for other single factor-dependent variable effects while effect sizes under the interaction of response model and ESV location are all very small.

4.3.2.2. Contrasts

Contrast tests are designed to assess the differences between dependent variables within each factor level group. Results for contrasts defined on response model and vehicle location are shown in Tables 4.5 and 4.6.

Table 4.5. Contrasts: Response Model

Dependent Variable	Difference*	95% CI for Difference	p
P1 Response Time	-3.60	-3.91; -3.21	< 0.001
P2 Response Time	-1.27	-1.95; -0.60	< 0.001
% P1 Meeting Target	2.87	2.27; 3.48	< 0.001
% P2 Meeting Target	0.44	0.01; 0.86	0.044

*(Single-tier) – (two-tier)

The greatest difference between response model factor levels was seen in P1 response times, followed by P1 responses meeting the response target. Differences involving P2 responses were less pronounced. All of the differences identified were significant.

Table 4.6. Contrasts: Vehicle Location

Dependent Variable	Difference*	95% CI for Difference	P
P1 Response Time	-7.37	-7.68; -7.06	< 0.001
P2 Response Time	-14.39	-15.07; -13.71	< 0.001
% P1 Meeting Target	28.55	27.94; 29.15	< 0.001
% P2 Meeting Target	4.71	4.29; 5.32	< 0.001

*Dynamic - static

Differences between levels of the factor ESV location above were generally of a greater magnitude than those for the response model factor. Summed absolute differences for response model were 8.187 vs. 55.02 for ESV location, which agrees with the importance of the ESV location factor to the model as identified in 4.3.2.1 above. Again, all differences were significant.

4.3.2.3. Interaction Plots

Interaction plots, of the estimated means of each dependent variable across levels of each factor, are shown in Figs 4.3 - 4.6.

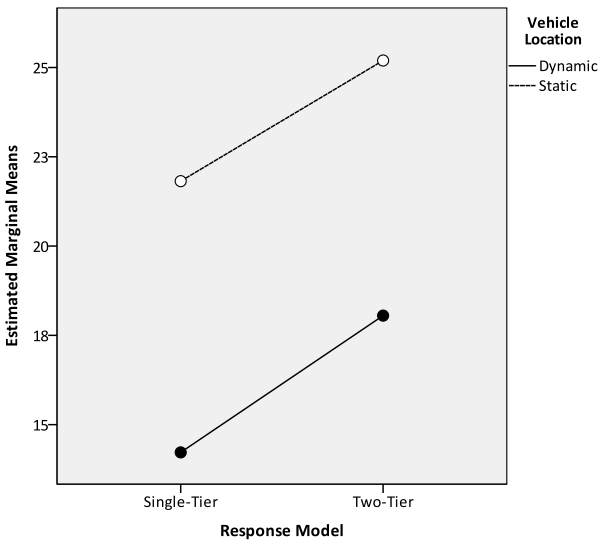


Figure 4.3. P1 Response Times

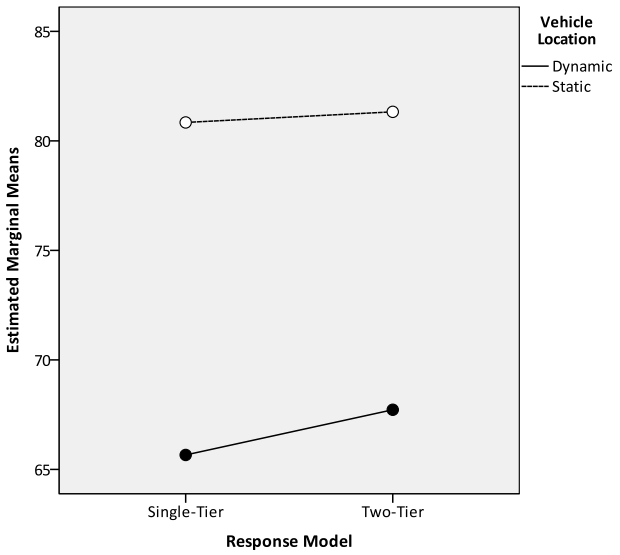


Figure 4.4. P2 Response Times

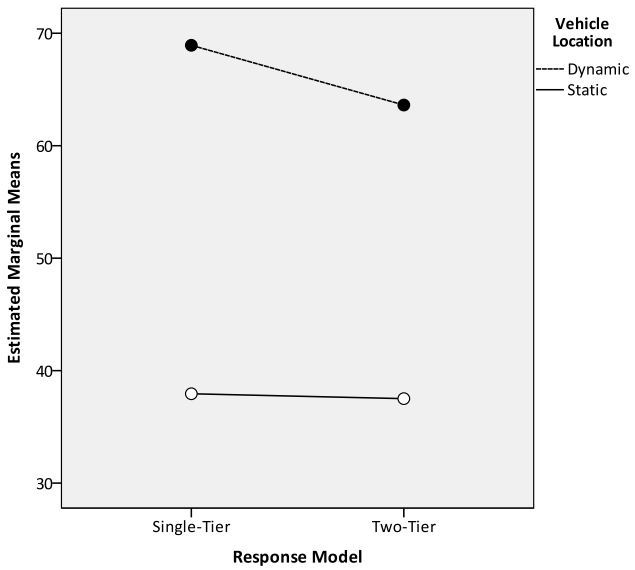


Figure 4.5. P1 Responses Meeting Target

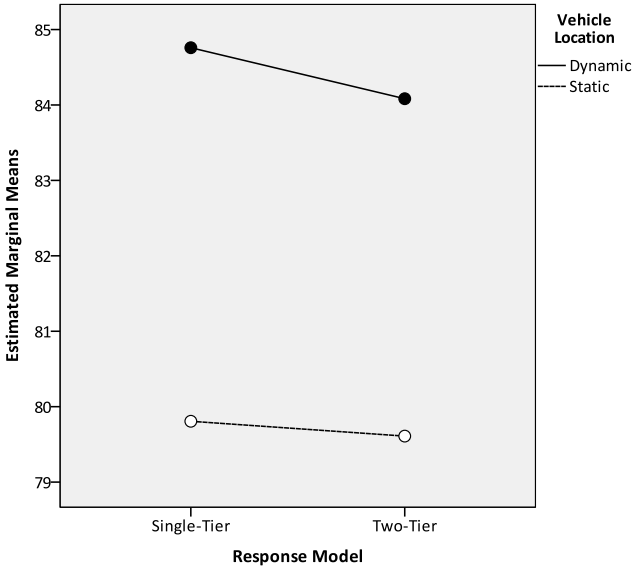


Figure 4.6. P2 Responses Meeting Target

The above plots confirm visually what the results in Tables 4.4 - 4.6 indicate, namely that dynamic vehicle location is associated with the best response time performance in both the P1 and P2 groups. The impact of the two different response models can also be seen, however this effect is less pronounced in general. The choice of response model does have more of an impact on response times across levels of vehicle location for P1 cases, than for P2s (Fig. 4.3 vs. Fig 4.4). This effect is not evident however when considering the proportion of P1 and P2 responses meeting response time targets (Figs 4.5 vs. Fig 4.6). Although the differential slopes of lines in Fig 4.5 and 4.6 confirm the interaction effect (Table 4.4), contrasts show that both individual main effects are significant for P1 and P2 cases (Tables 4.5 and 4.6).

4.4. RESEARCH OBJECTIVE 3: VEHICLE NUMBERS AND RESPONSE TIME PERFORMANCE GOALS

The last research objective examined the effect of vehicle numbers on response performance. As previously stated, a single best-performing model was to be selected based upon response performance - this is the single-tier dynamic model as identified from analysis in 4.3.2.1 and 4.3.2.2 above. In the event that this model was not associated with response performance meeting national benchmarks, ESVs were to be added to the model until this was achieved. As indicated above under 4.3.1, the single-tier dynamic model did not achieve national response benchmarks for P1 or P2 cases, and thus the conditional part of this objective was carried out.

Given the effect of vehicle location on response time performance identified above, the location of additional ESVs could be expected to exert a significant effect on response performance. The method used in deciding where to place additional ESVs was based on areas within each sector of greatest demand, as described in detail in 3.9.3.

4.4.1. Non-optimised Increase in Vehicle Numbers

Two approaches were followed in adding ESVs to the model. The first was an unrefined approach designed to obtain a crude idea of the upper limit of effectiveness associated with increased ESV numbers. In order to do this, ESV numbers were set to baseline values in each sector for the first scenario. After this, for each successive scenario, ESV numbers were increased by the original baseline number until either the response benchmarks were met or no further improvement was observed. Results of this process are shown in Table 4.7.

Table 4.7. Effect of Increased Vehicles Numbers on Response Targets

Scenario	Non-ALS	ALS	Total	Avail. (%)	P1 (%)	P2 (%)	%Diff. P1	%Diff. P2
1	38	15	53	54.9	68.93	84.76	-	-
2	76	30	106	78.2	79.50	86.17	14.24%	1.65%
3	114	45	159	85.6	81.61	86.23	2.62%	0.07%
4	152	60	212	89.2	82.86	86.25	1.52%	0.02%
5	190	75	265	91.4	83.40	86.19	0.65%	0.07%
6	228	90	318	92.8	83.94	86.27	0.65%	0.09%
7	266	105	371	93.9	83.90	86.27	0.05%	0.00%

P1 = Priority 1, P2 = Priority 2, ALS = Advanced Life Support, Avail. = Availability, Diff. = Difference

The initial doubling in ESV numbers produced a greater increase in P1 cases meeting the response time target than P2 cases, with these gains rapidly decreasing in scenario 2 and subsequent scenarios. Increasing ESV numbers to seven times their baseline value produced a 14.97% increase in P1 cases meeting the response time target while the same ESV number increases brought about only a 1.51% comparative increase for P2 cases. ESV availability increased with each successive increase in ESVs, however this increase was greatest with scenario 2's increase in ESVs and rapidly decreased thereafter. The total increase in ESV availability required to maximise the proportion of P1 and P2 cases meeting their respective response time targets was 39.0%.

4.4.2. Optimisation

When ESV numbers were increased as described above, a crude approach was taken of simply adding the baseline values for each ESV group in each sector to the previous values. No attempt was made to examine the effects of different combinations of ESV numbers across sectors or vehicle types (ALS vs. non-ALS) and how this would affect response performance.

In order to obtain an impression of which combination of ESVs would maximally increase P1 and P2 response performance optimisation was used, as described under 3.10.2 and 3.10.3 in Chapter 3. In addition to the settings described in 3.10.3, an upper limit of ESV numbers as shown in Scenario 7 of Table 4.7 was used together with a limit of 50 scenarios. Results of the optimisation process are shown in Table 4.8, with the last row of Table 4.7 reproduced for comparison (P1 and P2 proportions meeting the response target in the last row of Table 4.8 are cumulative). Cells with arrows indicate the direction of change from baseline (i.e. scenario 1) numbers.

Table 4.8. Vehicle Numbers and Response Performance: Non-optimised vs. Optimised

Scenario	ALS		Non-ALS		Total		P1 (%)	P2 (%)	%Diff P1*	%Diff P2*
1	15		38		53		68.93	84.76	-	-
17	105	90 ↑	226	188 ↑	331	278 ↑	84.0	86.4	19.7%	1.9%
7	105	90 ↑	266	228 ↑	371	318 ↑	83.9	86.3	19.6%	1.8%

P1 = priority 1, P2 = priority 2, ALS = Advanced Life Support, Diff = difference, * Proportional differences are given cumulatively, based on data presented in Table 4.7

Optimisation identified one scenario (number 17) that marginally increased both P1 and P2 response performance. More importantly, this was done with 40 fewer non-ALS ambulances than was possible with the non-optimised approach.

4.5. THE DISPATCH HAND-OFF DELAY

Results obtained from adding ESV numbers to the single-tier static model in order to assess what effect this would have on response performance have been presented in 4.4. A noticeable feature of the increases in P1 and P2 response performance are that they diminish rapidly after the first set of increased ESV numbers.

In order to identify the possible cause of this pattern, the delay between completion of the dispatch process and the allocation of a vehicle to a given incident was recorded in the model, referred to in this study as the *hand-off delay* (i.e. the delay in handing an incident off to a vehicle for response once the dispatcher is ready to do so). Any delay at this stage will add to the total response time. Additionally, it seems plausible that increasing ESV numbers may decrease this delay as greater ESV availability would make it more likely for a ESV to be available when any given incident is ready to be allocated. The nature of this delay was thus a likely explanatory candidate for the effect described above. Hand-off delay data are shown for plots of P1 response time (Fig 4.7) and proportion of P1 cases meeting the response target (Fig 4.8).

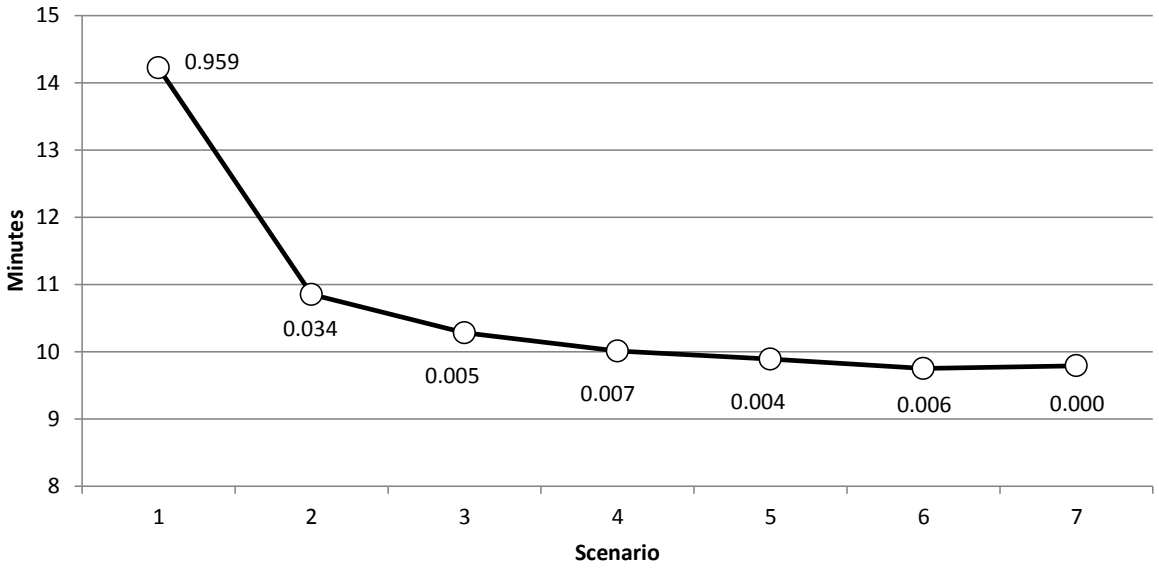


Figure 4.7. P1 Response Times Across Scenarios Showing the Hand-off Delay

Both plots (Figs 4.7 and 4.8) confirm that changes in P1 response time and the proportion of P1 cases meeting the response target are of a similar relative magnitude at each value of hand-off delay between scenarios 1 and 2. The initial, and largest, 186% negative difference in hand-off delay seen between scenarios 1 and 2 is associated with the largest increase (14%) in P1 cases meeting the response target (Table 4.7). This is associated with a decrease in response time of 27% in the same interval (between scenarios 1 and 2). In both cases, gains in response performance diminish with decreasing hand-off delays, giving the flattened curves of Figs 4.7 and 4.8. This suggests that once the hand-off delay has been eliminated, increased vehicle numbers offer no advantage in response performance.

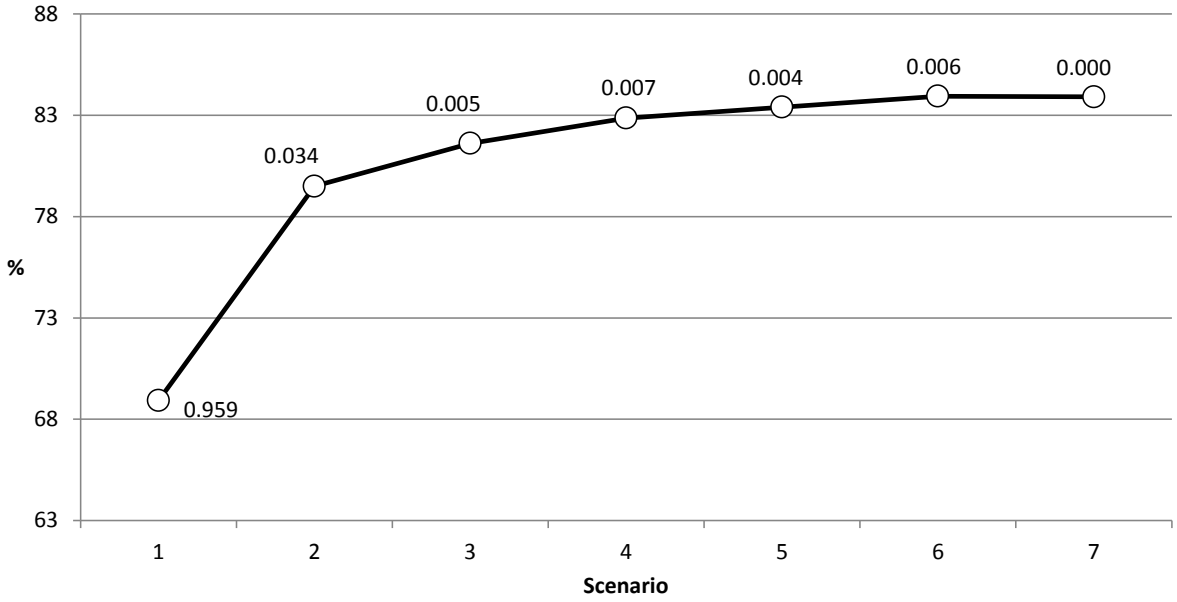


Figure 4.8. P1 Response Target Compliance Across Scenarios Showing Hand-off Delay

4.6. VEHICLE AVAILABILITY

Considering the effect of hand-off delay on response performance, and the impact of ESV numbers on hand-off delay, ESV availability data were recorded during each simulation run for each different model in order to assess the effect of response model and ESV location on this variable. ESV availability in this context is defined as the number of available ESVs remaining, out of the total number in the system. Fig 4.9 (on the next page) shows a typical time plot for ESV availability taken from one replication of the single-tier dynamic model. The mean of these observations over the whole replication was calculated, and the mean of these values over 15 replications was used to produce the summary ESV availability data for each combination of experimental factors shown in Table 4.9. Response performance rankings are shown in the same table for comparison.

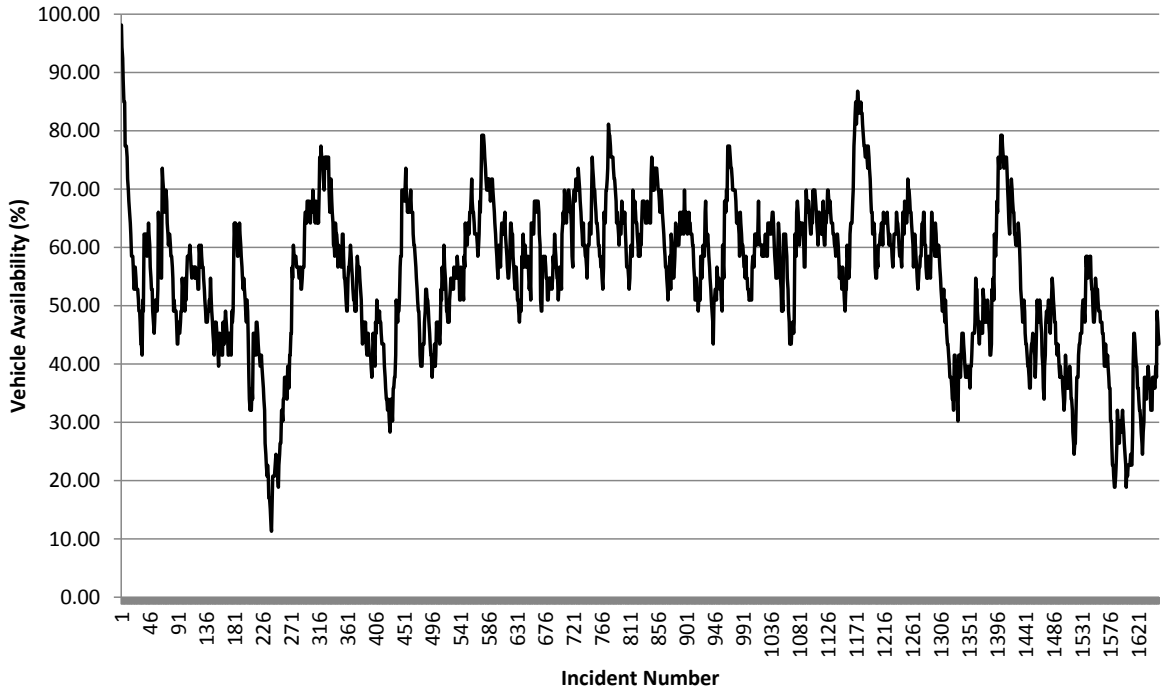


Figure 4.9. Typical Vehicle Availability Over Seven-day Period

Table 4.9. Descriptive Data: Vehicle Availability

Response Model	Vehicle Location	Response Performance Ranking	Mean (%)	95% CI
Single-tier	Dynamic	1	54.9	54.27; 55.50
Single-tier	Static	3	49.7	48.92; 50.40
Two-tier	Dynamic	2	45.8	45.14; 46.48
Two-tier	Static	4	45.3	44.54; 46.00

95% CI = 95% Confidence Interval

Descriptive data indicate that, within the ESV location factor, vehicle availability was greater for single-tier response models, which in turn had higher response performance rankings. Univariate analysis of variance was used in order to assess the effect of response model and ESV location on ESV availability. Fit of the statistical model was suitable with analysis of residuals demonstrating zero mean, constant variance and normality (Kolmogorov-Smirnov test, $p > 0.05$). Some evidence of negative autocorrelation was found in the residual data, however this was unlikely to have influenced conclusions drawn from the results in Table 4.10.

Both factors had a significant effect on vehicle availability, however the effect of response model was greater than that of vehicle location (Table 4.10, partial η^2 0.887 vs. 0.591). The interaction effect was also significant, however it was smaller than either of the factors alone.

Table 4.10. Univariate Test Results

Effect	F	df	p	*R^2	Partial η^2
Corrected Model	191.372	3	<0.001	0.906	0.911
Response Model	440.335	1	<0.001	-	0.887
Vehicle Location	80.856	1	<0.001	-	0.591
Interaction	52.925	1	<0.001	-	0.486

* Adjusted, Interaction = Response Model*Vehicle Location, df = Degrees of Freedom

In keeping with the effect size differences above, contrasts identified a 2.3 times greater difference for vehicle availability between levels of the response model factor than for those of the vehicle location factor (Table 4.11).

Table 4.11. Vehicle Availability: Contrasts

Factor	Difference*	95% CI for Difference	p
Response Model	6.74	6.09; 7.38	< 0.001
Vehicle Location	2.87	2.24; 3.53	< 0.001

*Difference between factor levels for vehicle availability

The interaction plot for vehicle availability estimated marginal means (Fig 4.10) shows that availability is lower in the two-tier models than the single-tier models for both levels of vehicle location, and that dynamic vehicle location is associated with greater vehicle availability across both levels of response model.

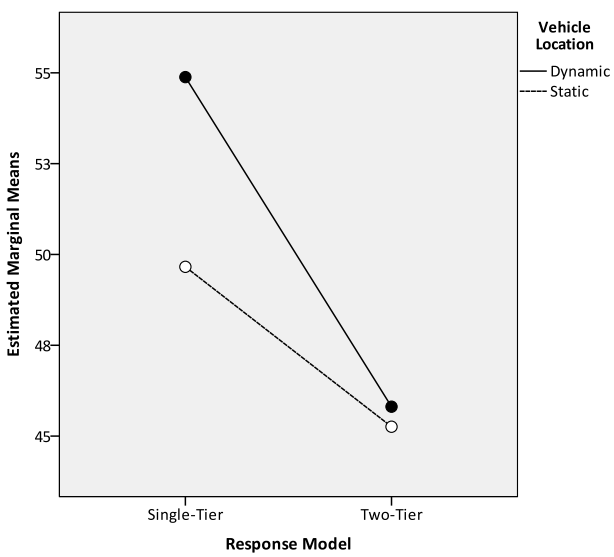


Figure 4.10. Vehicle Availability

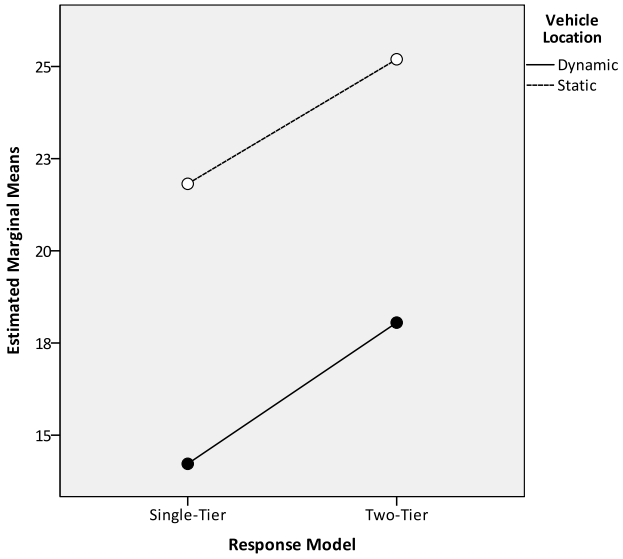


Figure 4.11. P1 Response Time

An ordinal interaction between response model and vehicle location is evident from the different line slopes in Fig 4.10, meaning that the availability values for response model are dependent on the availability values for vehicle location. Specifically, the effect of vehicle location appears to depend on the level of response model – the two-tier response model more negatively affects vehicle availability than the single tier response model for dynamic vehicle location. Despite the significance of this interaction (Table 4.10), contrasts indicate that both individual main effects are significant (Table 4.11).

When comparing Fig 4.10 and the interaction plot of P1 response times (Fig 4.11) an inverse relationship appears to be present between P1 response time and vehicle availability. This is in keeping with the observation that greater vehicle availability leads to a minimising of the hand-off delay and thus an overall reduction in total response time. The role of hand-off delay in influencing P1 response time is confirmed by mean P1 hand-off delays across the four different models as shown in Table 4.12.

Table 4.12. Hand-off Delay and Vehicle Availability Across Factor Levels

Response Model	Vehicle Location	Mean (min.)	95% CI
Single-tier	Dynamic	0.96	0.87; 1.05
Single-tier	Static	1.65	1.54; 1.76
Two-tier	Dynamic	1.95	1.84; 2.06
Two-tier	Static	2.65	2.53; 2.77

95% CI = 95% Confidence Interval

Greater vehicle availability appears to minimise the hand-off delay and this in turn results in shorter response times. Consequently, the model with the shortest hand-off delay (greatest availability) also has the shortest P1 response times (single-tier dynamic, Figs 4.10 and 4.11). The effect of vehicle availability on P2 response times is similar, but much smaller, because the P2 hand-off times were very small to begin with (ranging between 0.0002 and 0.0488 minutes).

4.7. OTHER OBSERVATIONS

Data presented in 4.3.2 above address the experimental objectives of this study. Vehicle availability, although not explicitly identified as an independent variable in the objectives, is central to explaining at least part of the variation in response system performance and is thus dealt with individually in 4.6. In the remainder of this Chapter other data obtained from model output are presented. These data have been included because they offer another perspective on the suitability of EMS system response (in addition to the sole consideration so far of response time) and because they will be linked to discussion of the above results in the next Chapter.

4.7.1. Appropriateness of Care

In the context of this study appropriateness of care refers to the match between patient acuity (as defined by patient priority) and the level of care delivered by practitioners dispatched to an incident of a given acuity. As explained in 3.4.2.4 of Chapter 3 on model process logic, the general aim of prioritised dispatch is to attempt to maximise this match over all incidents in a given time frame, and to ensure that ALS-level resources are dispatched to as may P1 incidents as possible. Although the major emphasis in this study has been placed on the investigation of response time as measure of system quality, appropriateness of care could be considered another important index of quality. Two variables offering perspectives on the appropriateness of care from model data obtained during experimentation are given below.

4.7.1.1. Advanced Life Support for Priority 1 Responses

This variable is a proportional measure of the matching of patient acuity and resources referred to above. Data were obtained by dividing the number of P1 incidents where ALS vehicles (either an ALS ambulance or a PRV) arrived at the incident node with the count of all P1 incidents per replication (means are calculated over 15 replications) (Table 4.13). Included in Table 4.13 is an additional measure of ALS involvement for P1 cases applicable to two-tier models only, namely whether a PRV arrived at the incident node first.

Table 4.13. Advanced Life Support On Scene

Response Model	Vehicle Location	Criterion	Mean (%)	95% CI
Single-tier	Dynamic	ALS at Incident	53.7	53.40; 54.01
Single-tier	Static	ALS at Incident	67.6	67.19; 68.05
Two-tier	Dynamic	ALS at Incident	48.8	48.43; 49.15
		PRV First at Incident	4.3	4.08; 4.43
Two-tier	Static	ALS at Incident	55.5	55.18; 55.86
		PRV First at Incident	4.9	4.74; 5.13

95% CI = 95% Confidence Interval, ALS = Advanced Life Support, PRV = Primary Response Vehicle

ALS vehicles were present at P1 incidents in roughly half of all cases, with the single-tier static model providing the greatest proportion of P1s with ALS. In the remainder of cases, a non-ALS vehicle was dispatched either because no ALS vehicle was available (see 4.7.1.2) or because the non-ALS vehicle was closer and was thus expected to arrive at the incident before the ALS vehicle giving a better response time. In the two-tier models, a PRV (if available) was dispatched if the closest or only available ambulance was non-ALS (see Fig 3.7 in Chapter 3). As the data in Table 4.13 show, PRVs only arrived first at an incident in between roughly 4% and 5% of cases. In the remainder of cases where PRVs were dispatched, they would have arrived after the non-ALS ambulance or may have been cancelled if the non-ALS ambulance had left the incident node before arrival of the PRV.

4.7.1.2. Availability of Advanced Life Support for Priority 1 Responses

Availability of ALS for P1 responses was defined as whether there were any ALS vehicles available at the time of dispatch for every P1 incident in each sector. Data were obtained by counting the number of times the ALS vehicle capacity was equal to zero for P1 dispatches in a given sector and dividing this number by the total number of P1 dispatches. These proportions are shown in Table 4.14 across all factor levels.

Table 4.14. Availability of Advanced Life Support at Dispatch

Response Model	Vehicle Location	Mean (%)	95% CI
Single-tier	Dynamic	8.6	8.28; 8.84
Single-tier	Static	21.6	21.18; 22.03
Two-tier	Dynamic	28.4	27.79; 28.93
Two-tier	Static	40.7	40.09; 41.25

95% CI = 95% Confidence Interval

The proportion of P1 dispatches where ALS vehicles were unavailable had a wide range over factor levels, with a 130% difference between smallest and largest values. ALS vehicles were unavailable more often in two-tier models, however even the single-tier static model was associated with unavailability in close to one quarter of P1 dispatches. For single-tier models the lack of ALS vehicles at dispatch produced a one-to-one relationship with the presence of ALS vehicles at the incident node, because pre-emption was not a feature of these (or any of the models). This was different for two-tier models, where a PRV could still be dispatched to an incident after the initial (non-ALS) ambulance dispatch if it became available. Consequently, some two-tier P1 incidents may have been included in the proportions of both Table 4.13 and 4.14.

4.7.2. Response Distances

In order to gain a different perspective on response performance as represented by time intervals, distances between the location of vehicles at the time of dispatch and the patient entities that they had been dispatched to were measured and written to text files during simulation execution. Because the simulation models approximated vehicle movement along straight lines between locations and not over networks of paths or link, the distances measured were straight line distances between vehicle and patient. Mean response distance values in kilometres over 15 replications for each model and different vehicles and incident acuities are shown in Table 4.15.

Table 4.15. Response Distances

Response Model	Vehicle Location	Vehicle	Mean (km)	95% CI
Single-tier	Dynamic	P1 AMB	3.84	3.80; 3.89
		P2 AMB	3.90	3.86; 3.95
Single-tier	Static	P1 AMB	7.78	7.74; 7.82
		P2 AMB	7.86	7.82; 7.91
Two-tier	Dynamic	P1 AMB	4.39	4.33; 4.45
		P2 AMB	4.44	4.38; 4.49
		PRV	8.16	8.11; 8.22
Two-tier	Static	P1 AMB	8.00	7.97; 8.04
		P2 AMB	8.02	7.97; 8.05
		PRV	8.84	8.76; 8.92

95% CI = 95% Confidence Interval, P1 = Priority 1, P2 = Priority 2, AMB = Ambulance, PRV = Primary Response Vehicle

As expected, dynamic models were associated with smaller response distances because vehicle locations at holding points were decentralised and distributed over each sector's total area allowing closer proximity of vehicles to incidents. The response model also accounted for some variation in mean response distances, with single-tier models having slightly shorter responses than two-tier models.

4.7.3. Cross-sector Responses

Cross-sector responses are responses where a vehicle is dispatched from a sector other than the sector in which the incident is located. This occurs when no vehicles in a given sector are available, and another incident in that sector requires a response. All models were designed so that in cases like these a vehicle could be dispatched from a neighbouring sector if one was available. Data on cross-sector responses, shown in Table 4.16, were obtained by counting the number of times a vehicle was dispatched where the patient entity sector (the sector in which the patient entity was located) and vehicle sector identifier were different, and dividing these by the total number of dispatches.

Table 4.16. Proportion of Cross-sector Responses Across

Response Model	Vehicle Location	Mean (%)	95% CI
Single-tier	Dynamic	4.9	4.77; 5.04
Single-tier	Static	14.7	14.52; 14.97
Two-tier	Dynamic	8.1	7.95; 8.25
Two-tier	Static	15.7	15.52; 15.94

95% CI = 95% Confidence Interval

Dynamic models were associated with smaller proportions of cross-sector responses, implying that these models result in better local (i.e. within-sector) vehicle availability. Although it is considered ideal to keep all responses within a given sector, it is not necessarily the case that cross-sector responses result in longer response times as in some instances vehicle and incident locations may be in close proximity on adjacent sides of a sector boundary. In other cases, however, distances may be much greater, leading to longer response times.

4.7.4. Mission Times

Mission time is the time a vehicle is dedicated to a particular incident and not available for other work. This interval begins at the time the vehicle is dispatched and becomes mobile to an incident and ends when the vehicle is available again (see mission interval, Fig 1.1, Chapter 1). It therefore

excludes any time taken for activities in the dispatch centre. Mission times for ambulances (P1 and P2 responses) and PRVs across all factor levels are shown in Table 4.17.

Table 4.17. Mission Times

Response Model	Vehicle Location	Vehicle	Mean (min.)*
Single-tier	Dynamic	P1 AMB	96.05
		P2 AMB	101.17
Single-tier	Static	P1 AMB	102.91
		P2 AMB	116.26
Two-tier	Dynamic	P1 AMB	99.05
		P2 AMB	103.11
		PRV	36.75
Two-tier	Static	P1 AMB	105.72
		P2 AMB	116.56
		PRV	36.80

PRV = Primary Response Vehicle, *Ambulance values are summed from the separate intervals in Figs 4.1 and 4.2 (response travel + on-scene + transport + post-incident). PRV mission times were determined individually from simulation output data.

Mission time differences between P1 and P2 responses within factor level groupings are accounted for mainly by differences in travel response times, which are shorter for P1s. Some mission time differences are evident between static and dynamic models, however these are generally small while the response model appears to have a negligible effect on mission time. PRV mission times are much shorter than those of ambulances, reflecting the fact that in most cases PRVs do not accompany an ambulance from the same incident to hospital and are thus available in a much shorter time frame.

4.8. SUMMARY

Results presented in this Chapter indicated that both experimental factors, ESV location and response model, had significant effects on response time and proportional response time target compliance for P1 and P2 cases. Of the two factors, ESV location had the greater effect. Based on these response performance data, the single-tier dynamic model was identified as having the best response overall performance.

Addition of ESV numbers to the single-tier dynamic model, using both non-optimised and optimised approaches, yielded very little improvement in response performance for the large numbers of ESVs

involved. Data on ESV availability and the hand-off delay showed that the ESV number-related improvement in response performance for P1s was due to minimisation of the hand-off delay and that once this delay had been eliminated no further response performance improvement was possible.

CHAPTER 5: DISCUSSION

5.1. INTRODUCTION

This study set out to assess the effect of three factors on response performance in a large urban EMS system in South Africa using computer simulation. The study sought to answer the following question: *What is the optimal configuration of ESV location, response model and ESV numbers required to meet response time performance goals, or provide the best possible response performance, in a simulation model based on a large urban South African EMS system?* Given the uniqueness of South African pre-hospital care both in terms of the volume, acuity, location and nature of incidents and in terms of historical EMS system design approaches, it is important that the answer to this question took into account these characteristics as embodied in the simulation models used in this study.

5.2. FINDINGS

The detailed findings of this study have been presented in the results of Chapter 4, arranged under sections that relate to the objectives of the study. The purpose of this section is to synthesise the study's main findings in order to answer the research question. The research question, as stated in Chapter 1 and repeated above, is a composite containing questions about the effectiveness of three factors on response performance. In order to arrive at a clear answer to this question a number of smaller individual questions will be asked which, when considered together, will provide an answer to the research question as a whole.

5.2.1. Which Level of Vehicle Location Produced the Best Response Performance?

Dynamic vehicle location produced better P1 and P2 response performance. This is clearly seen by the significantly shorter response times and greater proportion of P1 and P2 responses meeting their respective response targets in both models using dynamic ESV location, compared to the models using static ESV location. The shorter response distances associated with dynamic ESV location suggest that the use of holding points as a form of decentralised ESV location simply improved the proximity of ESVs to incident locations, thus reducing response times. Movement of ESVs between holding points in response to finer-grained spatial changes in demand may have further enhanced the effectiveness of dynamic location.

5.2.2. Which Level of Response Model Produced the Best Response Performance?

The single-tier response model was associated with better P1 and P2 response performance, both in terms of significantly shorter response times and proportions of P1 and P2 responses meeting their respective response targets. Single-tier response models had a greater beneficial effect on response performance for P1 cases than for P2s, and produced greater differences for response times than for proportional response time target compliance.

5.2.3. Which Combination of Factors Produced the Best Response Performance?

The single-tier dynamic model was associated with the best overall response performance for P1 and P2 cases. Dynamic ESV location contributed a greater effect to this performance than did response model. This model did not, however, produce response time performance meeting the national benchmarks of ≤ 15 minutes for at least 90% of P1 cases and ≤ 60 minutes for all other cases.

5.2.4. What was the Optimal Number of Vehicles?

When this question is asked with reference to compliance with national response time benchmarks, the answer is that *no optimal number of ESVs was identified*. It is an important finding that, even without any predefined upper limit being placed on ESV numbers, compliance with these benchmarks was not attainable. When this question is asked in a less restrictive way, then the answer is that *a seven-fold increase in ESV numbers produced the best possible response performance* (although this would not necessarily be termed optimal).

5.3. FINANCIAL IMPLICATIONS OF INCREASING VEHICLE NUMBERS

The striking observation about data on ESV numbers and response performance is the large number of additional ESVs required to bring about very small improvements in performance. Even with these large ESV numbers the national benchmark thresholds for P1 and P2 responses described above were not met. These ESV numbers translate into unworkable cost implications for even the best-resourced EMS systems.

Using costing data obtained from the Western Cape EMS (De Vries S. Vehicle cost estimates. [online] Email to Shaheem De Vries 11 February 2014 [cited 14 March 2014]), the following itemised estimates were determined, per ESV:

Ambulance: *R550,000*

Equipment: *R413,206*

Staff: *R2,500,000*

The total cost per ESV is thus R3,463,206 (approximately US$324,346 at a R/US$ exchange rate of 1 US$ = R10.6775). This does not include ESV running costs (such as fuel) or the ESV future replacement tariff which is added to ESV costing in reality.

Taking the ESV numbers data presented in Table 4.8, a total of 278 additional ESVs were needed in order to realise a performance improvement of 19.6% for P1 and 1.8% for P2 responses in meeting their respective response targets. This translates to a total additional cost of R962,771,268, or expressed in cost per percentage of performance improvement, R49,120,983 for P1 cases and R534,872,926 for P2 cases. Considering that the greatest single gain in response performance (14.24% for P1s and 1.65% for P2s) occurred with the first increase of 53 ESVs in Table 4.7, it could be argued that this would be a more practical target to achieve. However even this relatively smaller number of ESVs is associated with a total cost of R183,549,918.

It is clear that the cost implications above are untenable. The expense of supplying the 278 additional ESVs referred to above is 1.5 times the 2012/2013 budget for emergency transport for the whole Western Cape Province. Even the smaller amount is equivalent to approximately 30% of the Province's budget. If this money were available, its use for this purpose would not be defensible because the national response performance benchmarks would still not have been achieved at this additional cost and would most likely never be, regardless of how many additional ESVs were placed in the system. Such vast expenditure on a system with a high level of inefficiency (as indicated by the ESV availability data in Table 4.7) which seems incapable of ever performing to the expected level would not make any sense.

5.4. THEORETICAL IMPLICATIONS OF THE STUDY FINDINGS

5.4.1. Vehicle Availability, Vehicle Numbers and the Hand-off Delay

Both ESV location and response model had significant effects on ESV availability, with dynamic ESV location and single-tier response models being associated with greater ESV availability. Of the two factors, response model was observed to have a larger effect on ESV availability although the relationship between ESV availability and the two experimental factors was modulated by an interaction effect. An association between hand-off delay and ESV availability was evident, although this did not appear to be linear.

The existence of a hand-off delay between the completion of dispatch and the allocation of an available ESV has been described in other simulation studies.(102,106) The effect of ESV numbers on this delay, and on response times, has also been studied and has similar characteristics to those found in the current study. In both cases, reduction of this delay by the addition of ESVs improved response times until elimination of the delay after which no response time improvement was possible. The ESV numbers required to reach this point in other studies were proportionally smaller than those seen in the current study. In the study by Savas, a 73% increase in ESV numbers cancelled the hand-off delay (106) while Inakawa, Furuta and Suzuki found that a 100% increase in ESV numbers had this effect.(102) This is in contrast to the 150% increase in ESV numbers required to cancel the hand-off delay in the current study. The reasons for this are not clear, however the higher incident rates and more complex dispatch and response systems in the current study may have increased the variance in response times and thus diluted the effect of additional ESVs, compared to the two smaller systems.

5.4.2. Vehicle Numbers and Availability as System Constraints

In attempting to understand the results from this study related to ESV numbers, ESV availability and response performance it is helpful to relate these to theory. One theory that has relevance in this type of system is Goldratt's theory of constraints (TOC), which is a systems-management philosophy focusing on constraints as the limiting factors in any organisation.(120,121) Although the TOC has been applied mainly to the areas of manufacturing, project management and supply chain solutions,(120) its use has extended to other areas of business and even to health care and emergency care operations.(122)

In TOC, a basic assumption is that processes or organisations function as chains. The emphasis in TOC is therefore on the balancing of flow rather than capacity. The TOC's fundamental assumption is that every system must have at least one constraint which limits it from improving its performance or achieving its goal. Usually there are a small number of constraints in a system, and these may be internal or external to the system, and either physical or managerial (in the form of policies, procedures, rules or methods). The identification of constraints, and the prioritisation of constraint exploitation are aimed at aligning all other parts of the system to support maximum effectiveness of the constraint. When steps aimed at the exploitation of constraints are effective, the constraint may be "broken", meaning that it is no longer the system's limiting factor.(120,121)

The single-tier dynamic simulation model used in this study can be considered as a system of the type Goldratt describes in his theory. In this view, ESV numbers can be identified as a possible system constraint as they affect ESV availability. From the data presented in Table 4.7 it is obvious that the ESV numbers required to reduce hand-off delay and thus improve response performance are substantial, and that these ESV numbers have only a small effect on improving response performance. Consequently, it is likely that ESV numbers are not the most important constraint on response performance, and in fact may be relatively less important constraints. Clearly, if the goal of a system such as that modelled in the current study, is to meet the national response benchmarks as defined earlier, then the number of ESVs and their resultant availability cannot be considered a constraint of any practical significance in achieving this goal. Several other constraints, both in the pre-dispatch and post-dispatch intervals, may be responsible for the impedance of response performance.

5.4.3. Queuing for Vehicles Despite High Vehicle Availability

Data presented in Table 4.7 and Fig 4.9 show the existence of a hand-off delay in the presence of ESV availability greater than 50%. Incidents were therefore queued, waiting for an available ESV, despite the fact that in the system as a whole at least half of the ESVs were available. Availability data in the last row of Table 4.7 indicate that close to 100% ESV availability was necessary in order to reduce the hand-off delay to a very small value. This result can be better understood by consideration of some fundamental principles of queuing theory.

A queuing system consists of one or more servers (objects that provide some kind of service), customers and a queue. The term “customers” is not meant literally, but is rather seen as any type of entity that requests some kind of service from a system. Queuing system are described by reference to three related components, namely the arrival process (describing how customers arrive to the system), the service mechanism (describing the number of servers, the number of queues and the probability distribution of service times) and the queue discipline (describing how the queue is ordered). A number of performance measures can be derived for queuing systems including the number of customers in the system and in the queue, the average time spent in the system and in the queue, and server utilisation. Several relationships between server utilisation and queue length are also described in the literature.(123–125)

The link between queuing theory and the ESV availability-hand-off delay relationships identified in Table 4.7 is that ESVs can be viewed as servers (i.e. they provide a transportation service to patient

entities) and are thus associated with a queue of patient entities. Time spent in this queue is equivalent to the hand-off delay. The known relationships between server utilisation and queue length may thus offer some explanation as to why a hand-off delay exists despite what appears to be relatively low server utilisation.

In the simplified case of a queue with an exponential arrival distribution, general service distribution and single server (a M/G/1 queue), queue length is dependent in part upon server utilisation. The greater the server utilisation, the longer the queue length tends to become.(123–125) In general, server utilisations (i.e. the complement of ESV availability) of the magnitude seen in Table 4.7 would not be expected to produce any queuing at all. One other variable however also affects queue length, and that is the variability in service time, specifically that the greater this variability is the greater queue length tends to become. Queue length may be decreased by decreasing server utilisation or by decreasing service time variability. Server utilisation in turn may be decreased by reducing the arrival rate, increasing the service rate or increasing the number of servers.(123–125)

The existence of a queue of patient entities in the presence of good ESV availability is thus most likely due to variation in transport times while variation in arrival times may also play a role. Unfortunately, little can be done in the short term to lessen the variability in either arrival rate or transport times. Increasing the service rate may be possible by reducing the mission time of ESVs, however the effect of this on hand-off delay is currently not known.

5.5. IMPLICATIONS FOR THE PROVISION OF ADVANCED LIFE SUPPORT

The matching of patient acuity with an appropriate level of care is an important goal of prioritised dispatch in systems where an ALS level of care is available. Although seldom mentioned in isolation as a performance indicator, many of the clinical performance indicators implicitly require ALS presence for high acuity cases because of the type of interventions associated with quality care.(41) The dispatch and ESV allocation logic used for both single- and two-tier systems in the current study placed emphasis on the matching of P1 incidents and ALS-level ESVs. However this was sometimes subjugated by the importance of response time, minimisation of which was the overarching objective.

Results presented in Chapter 4 indicated that an ALS ESV was present at incidents approximately half of the time across the different factor combinations. The higher rate of ALS presence for the single-tier static model, compared to the single dynamic model, can be explained once again by the

proximity of ESVs to incidents. Because dynamic models increased this proximity, and because most of the ambulances in all sectors were non-ALS, this would have led to a situation much of the time where the closest ESV (which would subsequently have been allocated) was a non-ALS ambulance. In contrast, because proximity of all available ambulances to incidents in the static models was the same, ALS ambulances would have been dispatched more frequently as the closest ESV. Availability of ALS ESVs at the time of dispatch also contributed to the frequency of ALS presence at incidents, however this differed between single- and two-tier models as in the latter a lack of availability of PRVs at the time of dispatch did not necessarily preclude the eventual presence of these ESVs at a P1 incident, as explained in 4.7.1 of Chapter 4.

Of some significance is the lack of obviously better ALS attendance at incidents where a two-tier model was used. The presence of PRVs in these models, which are dedicated to only respond to P1 incidents, did not appear to increase ALS presence at incidents in any practically significant way. Once again, the difference in this respect between dynamic and static models was most likely the fact that dynamic models placed many non-ALS ambulances closer to incident locations thus making them prime candidates for allocation more often. Although PRVs were still allocated to P1 incidents when non-ALS ambulances were allocated, in many cases the PRVs may have been cancelled before arrival as a result of the longer distances they were required to travel on average for P1 responses.

Data on PRV responses to P1 incidents in the two-tier models indicate that these ESVs very rarely arrived first at the scene of an incident. Whether the ESV location model was dynamic or static did not appear to change this in any practically significant way. This effect is most likely a question of proximity, as a single PRV was assigned to cover each sector. This meant that response distances were typically long for PRVs and thus in most cases an ambulance arrived at the incident scene before a PRV. Although the argument could be made that the value of ALS was still realised even in cases where an ambulance arrived at the incident scene first, the basic premise of using PRVs in a system is to deliver ALS most effectively to incidents where it is needed. This would presumably mean that ALS should arrive at any P1 incident early, during the important patient assessment and initial resuscitation phase.

The above arguments in relation to the adequacy of ALS provision by different factor models are offset at least partially by considerations other than those centred solely on response performance. In reality, not every P1 case requires the presence of ALS at an incident for the provision of an adequate standard of care (as an example, the unconscious hypoglycaemic patient requiring

intravenous glucose which can be administered by non-ALS personnel). However accurate dispatch triage is required in order to correctly filter such cases, which will still represent a minority of incidents. On the other hand, in South Africa at the present time, analgesia can only be administered by ALS-qualified personnel meaning that ALS might be required for this purpose even at incidents that may not otherwise fit the criteria for classification as P1, further heightening the importance of the system's ability to provide this resource.

Retention of ALS personnel over time is another consideration impacting the arguments made above, which suggest that single-tier models are preferable. In two-tier systems as modelled in this study, PRVs staffed exclusively by ALS personnel are only dispatched to P1 incidents. Consequently, ALS personnel in this kind of system could be rotated through duty on the PRV as well as duty on ALS ambulances. In single-tier systems there are no PRVs and thus ALS personnel would only be allocated to ALS ambulances which may spend some of their time servicing lower acuity (P2) incidents (as they would also in two-tier models). This may be perceived by ALS personnel as an inappropriate use of their knowledge and skills leading to frustration and, possibly over time, a greater probability of their intent to leave the clinical practice environment.

Although the argument above bears some anecdotal truth, there is little quality evidence to support it. Few studies have assessed the perceived importance of dealing with high acuity cases on ALS personnel and not in enough detail to consider the results as directly applicable to the current context. In the Longitudinal Emergency Medical Technician Attribute and Demographic Study a high proportion (89.7%) of US Emergency Medical Technician Paramedics (EMT-Ps) rated "having a job that is exciting" as moderately or very important.(126) Intent to leave the clinical environment was not, however, assessed in this study. Another US study found that EMT-Ps having fewer emergency calls and more scheduled transfers was a factor related to a significantly greater intent to leave the clinical practice environment.(127) On the other hand, in a South African study assessing reasons for withdrawal from clinical practice amongst a group of ALS-qualified personnel who had already done so, exposure to high vs. low acuity incidents was not identified as a factor influencing this decision.(128) The majority of respondents in this study most likely did work in two-tier EMS systems, thus it is not possible to infer from these data what the effect of changing this part of the work environment would have.

Although results of the above studies were based on fairly non-focused questions (relating to an "exciting job" or comparing scheduled transfers with emergency calls), their results suggest that

there may be some risk in diluting the expected high acuity work load of ALS-personnel with lower acuity incidents, and that this may result in some lack of retention over time. More research is needed in order to establish how significant this risk is, and if there are any other factors related to job satisfaction which may mitigate it. In the longer term, perhaps a perceptual shift is required in the South African pre-hospital emergency care education and operational domains that can more closely align the reality of urban EMS operations with the expectations and clinical skill set of those working in this area.

5.6. IMPLICATIONS FOR THE DESIGN OF EMERGENCY MEDICAL SERVICES RESPONSE SYSTEMS IN SOUTH AFRICA

Results obtained from the four different simulation models in this study provide useful information about the relative importance of the chosen experimental factors in altering the effectiveness of response performance in a simulation model based on a large urban EMS in South Africa. Perhaps the single most significant observation arising from this study is that the manipulation of ESV location, response model and ESV numbers in various ways did not bring about a situation where response performance of the simulated system met the national benchmark of 90% of P1 responses within 15 minutes and all other responses within 60 minutes. This provides useful information on the type of system characteristics which could be expected to bring about better response performance, but also where the placement of resources is likely to be futile and thus avoided. Some implications of this study's results are discussed below.

5.6.1. The Primacy of Vehicle Location in Determining Response Performance

Of the two experimental factors investigated in this study, ESV location was the most important with regard to an effect on response performance. The use of holding points in order to improve the proximity of ESVs to incidents stood out as the single most important factor in reducing response times, especially those for high acuity cases. Dynamic ESV location was associated with the shortest response distances, the greatest ESV availability, the smallest hand-off delays, the greatest availability of ALS ESVs at dispatch, the smallest proportion of cross-sector responses and the shortest mission times.

Holding points included in the baseline simulation model (and in the experimental models that used dynamic ESV location) were close approximations to those used in the real system. The exact method used for determining the location of these points in the real system was not entirely clear, other than that the points were considered to be central to areas of high incident density based on

historical data and experience of the system. The policy of moving available ESVs into a holding point with zero capacity was also modelled in order to simulate real ESV movements as they occurred in the system. The principle of placing and re-allocating ESVs according to approximations of demand was thus shown to be effective in reducing response times in the simulation model, and this appeared to be achieved mainly by minimising response distances.

Pioneering work performed by the RAND Institute's Fire Project in the 1970s in order to study the relationship between varying demand for firefighting resources, deployment of these resources and response distance (and hence time) established an important relationship between response distance and the number of response locations in a given area.(129,130) This work showed, in a fire suppression context, that the average response distance in a region is inversely proportional to the square root of the number of response locations, equivalent to holding points in the current study if the assumption is made that the holding point locations are not changed in the short term. A model developed as part of this work allowed prediction of average response times based on parameters of the coverage area, the number of response locations, the arrival rate and distribution of calls (the same as incidents), the mission time and a proportionality constant based on geographical relationships and characteristics in the model.(129)

The "square root law", as it has become known, underscores the importance of decentralisation in placing resources closer to finer-grained areas of deployment, and thus reducing response distance and time. Although the current study did not make any comparisons in this regard beyond a crude centralised (static) vs. decentralised (dynamic) contrast, the obviously better response performance of a decentralised approach supports Kolesar and Blum's distance-location relationship (129) and suggests that response performance improvement may be further enhanced by more decentralisation of ESVs. Many related questions require clarification, however, including the effects of local geo-spatial relationships and in particular the effect of informal settlements which are an ever present feature in South African urban centres. The question of ESV numbers and their effect on response performance, answered in the current study in relation to the specific arrangement of holding points and incident location patterns, would be of renewed relevance in any study of greater decentralisation as this would most likely change the fundamental dynamics of the model.

5.6.2. Intelligent Vehicle Location: Anticipating Demand

In the simulation model used for this study, a decentralised ESV location using holding points was modelled. However the positioning of holding points in areas of high demand appeared to be

somewhat loosely based on historical data. A number of alternative approaches are possible that have the potential to statically deploy ESVs in an optimal way based upon a range of mathematical and analytical techniques.(90,97–100,104) Similarly, variance in system demand can be incorporated into both heuristic and analytical methods for determining optimal redeployment policies such that ESVs are located in anticipation of demand, thus minimising response distance and time.(88,93–95,109)

Although many of these approaches are technically complex and may be associated with considerable development time and cost, the potential for real response performance improvement is likely greater than would be the case if resources are simply increased without careful consideration of their deployment and redeployment. As discussed in 5.6.4 below, simply increasing ESV numbers in a system without an optimised deployment and redeployment policy will amount to a waste of these resources.

A fundamental assumption of the observations above (in 5.6.1 and 5.6.2) is that a reasonably developed EMS systems foundation is available for further development. This would necessitate a dispatch centre with a functional CAD system, good communications infrastructure and a fleet of serviceable ESVs that is at least in line with planned targets for a given system, however these are derived. The current study investigated response performance using a model of a large, urban EMS and it is in this context that the system enhancements concentrated on decentralisation and demand-based ESV deployment and redeployment could potentially lead to response performance improvements. It is also within such centres, the four largest being the cities of Cape Town, Johannesburg, Pretoria and Durban, that the best resourced and most foundationally sound EMS systems are expected to be found. EMS systems in smaller peri-urban centres or rural areas, which are generally less well-resourced and developed, require separate analysis and modelling in order to determine where structural and process adjustments would be likely to have the greatest impact on response performance.

5.6.3. Transport vs. Non-Transport Advanced Life Support Vehicles

As described in 1.2.3.2 of Chapter 1, the use of PRVs as the first level of a two-tier design in urban South African EMS systems is commonplace. The reasons for this are centred mostly on what has been perceived historically to be the best use of scarce ALS-qualified personnel. The use of PRVs is believed to yield the best area coverage of ALS personnel and to allow good response performance for P1 incidents.

The results of this study contradict the rationale for PRV use, both in terms of response times and in terms of provision of ALS-qualified personnel. Two-tier models were associated with longer P1 response times, longer P1 response distances, proportionally less P1 responses within the 15 minute target, lower availability and greater hand-off delays. In the case of static ESV location, the two-tier response model did result in slightly greater ALS provision at incidents. However, when considering that PRVs were the first to arrive on scene only 4.9% of the time the premise of PRVs being responsible for the provision of early ALS at incidents is questionable. Given the fact that the dynamic two-tier model used in this study located PRVs and ambulances (both ALS and non-ALS) differently, with PRVs on average covering response distances at least twice those of ambulances, it becomes clear that the response performance of a few PRVs cannot match that of a larger number of ambulances located at holding points, with the proximity advantage that this offers.

5.6.4. The Limited Benefit of Many Vehicles

This study demonstrates the very limited effect of increasing ESV numbers in an urban EMS system in isolation. The number of ESVs required to bring about any response improvement of practical significance, and the associated cost, is clearly beyond the reach of any EMS in South Africa. More pertinently, the consideration of such expenditure even if it were available would be completely illogical and wasteful. When EMS decision-makers are faced with choices regarding expenditure, they should consider how their existing resources are being used before deciding to spend more on increasing these in the hope of improving response performance. In particular, ESV deployment and redeployment policies are most likely the area of greatest return on investment for post-dispatch response performance.

Quite aside from simulated data presented in this study, the real situation in KwaZulu-Natal Province, as explained in that Provincial Department of Health's most recent Annual Report, seems to support the observations arising from this study with regard to ESV numbers . In the report, the Department points out that "…*increase of vehicles to improve efficiencies seems not to have the expected results*."(33) This is in reference to the addition of 310 ESVs to the Province's fleet in 2012/2013, with a coincident decrease in response performance. Although the Gauteng Province's Annual Report does show an improvement in P1 response performance of around 30%, only 28 ESVs were apparently added to the fleet. Repeated reference is made in this report to the objective of procuring more ESVs in order to enhance response performance.(32) Although the data above are Provincial and not from urban centres alone, it appears that the supposed relationship between

response performance and ESV numbers is firmly entrenched although there is no real evidence to support it.

Even though the limited effect on response performance of adding ESVs to an urban EMS system has been highlighted, it would be incorrect to assume that the number of ESVs in such a system has no bearing on response performance at all. Despite not having been investigated in this study, it is logical to believe that as ESV numbers in a system are decreased the hand-off delay will increase and thus response performance will be adversely affected. The precise nature of this relationship is however not known and worthy of further study.

From a managerial point of view, it would be beneficial for decision-makers to know what the required minimum number of ESVs for their EMS system is. However this apparently simple question is fraught with difficulty. It is unlikely that the question of optimal ESV numbers can ever be considered apart from the question of how those ESVs are deployed and redeployed within the very specific demand and spatial characteristics of any given EMS system. Thus the question must be answered in every unique context. In the current study, the problem of determining an optimal number of ESVs in order to comply with national P1 and P2 response performance benchmarks was unsolvable because the deployment and redeployment policies applied did not allow these benchmarks to be reached, regardless of how many ESVs were added to the system.

5.6.5. The Need for Formal Assessment of Emergency Medical Services System Constraints

As the discussion in 5.4.2 above has emphasised, improvement of the response performance of EMS systems depends on knowing what the system constraints are. In the current study, ESV numbers and the related variable of ESV availability were assumed to be possible constraints on response performance. Although this is true, results clearly showed that other system constraints must be responsible for the inability of this simulated system to meet its goal of compliance with the national P1 and P2 response time benchmarks. There is thus a need for a more formal approach to the study of EMS system processes, and in particular constraints, in order to determine where effort and resources are best spent in improving response performance.

5.6.6. The Area of Greatest Immediate Gain: The Dispatch Delay

Objectives of the current study focused only on assessing the effect of ESV location, response model and ESV numbers on response performance after the dispatch process. In other words, the choice of experimental factors assumed that processes after those occurring in the dispatch centre could be

modulated in order to significantly improve response performance. Consequently, dispatch processes were simply modelled on input data from the real system.

The input data on delays accounting for call taking and ESV dispatch suggest that dispatch processes in the modelled system would be good candidates for constraint analysis. The baseline simulation model in the current study simplified these processes down to a single process with an associated delay, and no doubt the real underlying processes are much more complex. The current study also did not factor the influence of inter-hospital transfers, which likely consume a significant amount of dispatch centre capacity, into dispatch centre activities and the effect that this may have on the efficiency of dispatch for primary response cases. Nevertheless, shortening of the delays associated with P1 and P2 call taking and dispatch could make a significant contribution to improving overall response performance even before the complexities of ESV deployment are tackled. Simulation may also be a useful tool for better understanding and comparing possible approaches in dispatch processing.

5.6.7. Validity of the National Response Time Benchmark

Origins of the benchmarks used for response performance as cited by Provincial Departments of Health are uncertain.(31–33) There does not appear to be any acknowledged evidence upon which to base these choices of thresholds for “adequate” performance. In this sense, the South African position is not much different from that in the US and UK, except that the “eight minute rule” discussed in Chapters 1 and 2 is based upon some scientific evidence, even if the generalisation of this from a small set of cardiac arrest cases to all P1 incidents is misplaced.

This study has shown that even with a decentralised ESV location strategy and a dispatch centre capable of processing P1 cases in a relatively short time period, the response time benchmarks were simply unattainable for either P1 or P2 cases, regardless of the number of ESVs in the system. It is possible that by reducing delays in the dispatch centre even further (as suggested in 5.6.6) the P1 and P2 benchmark thresholds for response time might be reached. Beyond this, response times will most likely only be improved by further decentralisation and deployment strategies taking shifting demand patterns into account. At the present time, it is unlikely that any urban EMS in South Africa is in a position to implement such an approach successfully.

5.6.8. Implications of the Study Findings in Context

The research question in this study was motivated primarily by many years of experience of poor response performance in an urban setting. These observations are supported by data reported year after years in annual reports showing an inability of urban EMS systems to meet response time benchmarks that are, by international standards, quite conservative. In attempting to understand why urban EMS systems in South Africa have evolved this way, it is helpful to briefly consider possible factors influencing this evolutionary process and to relate these to the study findings.

The Origins of Poor Response Performance

There has been a long history of centralised ESV location in urban EMS systems in South Africa. This has presumably been brought about by the early influence of Fire Departments in the provision of EMS, and the typical Fire Department-orientated approach to vehicle deployment using a small number of fixed locations serving relatively large areas. Even though Provincial Departments of Health are now directly responsible for running EMS in most Provinces in South Africa, for the most part a dominantly centralised approach to ESV location has persisted. Additionally, locations chosen as bases from which ESVs operate typically are existing structures owned by Health or Fire Departments (e.g. clinics, hospitals or fire stations) and have not necessarily been selected because they have close proximity to areas of incident density and demand. Thus the typical approach to ESV location in urban South African EMS systems has evolved as one most closely resembling the static ESV location level represented in the simulation models used in this study.

In common with EMS development in many other countries, qualifications and scope of practice in South African EMS in the 1970s were what would today be referred to as a BLS level of care. Following developments mainly in North America, South African EMS training and qualification levels were expanded in the 1980s to include ILS and eventually ALS levels of care. From the outset, the number of emergency care personnel with ALS qualifications was relatively small – a trend that persists today. Perhaps for this reason, with the advent of ALS qualifications a two-tier approach to EMS response was initiated with ALS qualified personnel responding to P1 cases in PRVs and ambulances staffed by non-ALS qualified personnel responsible for transporting these cases to hospital. The thinking behind this approach was most likely to enhance the availability of ALS by placing them in non-transport ESVs. Thus the typical response model in urban South African EMS systems has evolved as one closely resembling the two-tier response model represented in the simulation models used in this study.

Considering the way that EMS response has developed in urban areas in South Africa, and the characteristics that have evolved over time, it is not that surprising to see why these systems have failed to deliver acceptable response time performance. Most of these systems share characteristics with the worst-performing simulation models in this study, namely a predominantly static (and not demand-orientated) ESV location strategy and a two-tier response model. The findings presented in 5.2.1 and 5.2.2 clearly highlight the limitations of these features.

Data from annual reports of the Provincial Departments of Health suggest that under-resourcing, in the form of ESV numbers, is thought to be the cause of poor response time performance. This is a reasonable enough sounding idea, and ties in with a perception of poor health care delivery in South Africa generally being brought about by a lack of resources. However the findings of this study suggest that resources, in the form of ESV numbers, are not the cause of poor response performance, and consequently that increasing these will not address the problem.

A Mind-Set Change to Bring About Real Change

If the current national response time benchmarks are to ever be complied with in urban EMS systems, and if the matching of incident priorities with an appropriate level of care is important, a fundamental change of mind-set is required in the approach to designing urban South African EMS systems. Current consideration of EMS design in relation to response performance tends to focus on the perceived adequacy of the approaches used and lack of resources as a cause of poor performance. The focus on resources as the primary problem needs to change to a focus on the *adequacy of system design* in matching demand with resources, as the key driver of performance. In making this shift it may become apparent that there is no or little significant absolute deficiency in resources and that effective alignment of demand and resources may improve efficiency, something that was observed to be worryingly low in all of the simulation models.

The starting point for this change in focus is detailed knowledge of where demand in a system is, and how it changes in both the short- and long-term. For this to be realised, accurate incident location data are required along with methods to analyse these data. This is of such fundamental importance that it is doubtful that any significant improvement in response performance will be possible without it. Matching resource allocation with demand lies at the heart of planning an appropriate response and those responsible for this planning must understand the problem in order to solve it. This goes beyond the experience-based rough estimation of demand distribution and must be based on valid quantitative methods.

Beyond understanding the distribution of demand in an EMS system, all other efforts should be focused on proximity - in other words addressing the ESV deployment and re-deployment problems. This will necessitate fine-grained decentralisation of ESVs in order to minimise response distances, and anticipation of shifts in demand that allow for movement of ESVs in such a way that hand-off delays are minimised. The matching of P1 incidents with an appropriate (ALS) response will mean decentralisation and greater spread and number of ALS personnel and a move away from the principle of a small number of ALS personnel servicing P1 cases over a large area, which is a fundamentally flawed approach from a response performance perspective.

How Difficult Would This Change Be to Implement?

A natural and important question arising from consideration of the argument above is whether EMS personnel would be amenable to such a change in emphasis regarding day-to-day operations. At an operational level, decentralisation of ESVs and ALS capability would mean fundamental changes to the way many personnel typically operate today including more time spent in ESVs at holding points and for ALS personnel, a greater mix of P1 and P2 responses. Both of these would most likely be contentious issues and may be argued to have a possibly negative impact on retention of personnel, especially ALS. However the *status quo* regarding how EMS personnel operate in centralised systems, and the expectation that ALS personnel exist for the sole purpose of treating a small percentage of high-acuity patients, seem to be at odds with the design and operational requirements of systems characterised by agility, efficiency and optimal response performance. Deep-seated perceptions of the role of EMS personnel in the delivery of pre-hospital emergency care will need to be changed in the long term in order to address retention, and this can only be achieved through education and the articulation of a clear EMS strategy.

But What Would it Cost?

Obtaining a detailed picture of EMS system demand and its variation would require a significant investment in technology and skills considering that such technology and skills are only partially- or non-existent in most urban South African EMS systems. This, together with the modelling of decentralised systems and the decision-support technology that would make such systems work, has significant cost implications. A heavy reliance on technology, probably with a large bespoke component, also represents increased risk and this in turn has additional cost implications. However it is not at all clear that the current approach of trying to address flawed systems design with more and more resources, an approach which is very inefficient, would in the long run be more cost-

effective. Any solution that could more successfully match demand and available resources and thereby increase efficiency may represent an eventual cost saving with the appropriate managerial oversight and strategic control.

Other Questions Requiring Answers

Apart from the main shift in mind-set referred to above which is focused on the post-dispatch part of the response system, there are two other related question with the potential to significantly affect response performance, requiring answers. The first is that of dispatch efficiency and why the dispatch of incidents takes a long time, particularly for P2 incidents. The second is that of on-scene time and why this is so long. Shortening of both (or even one) of these intervals could have a significant effect in improving response performance and it thus makes little sense to place emphasis on other parts of the response system and not these. There may well be valid contextual reasons for the observed intervals involved in dispatch processing and on-scene times, however it is difficult to imagine that there is no possible way to make improvements in these areas at least to some extent.

5.7. LIMITATIONS

Several limitations must be kept in mind when considering the results of this study and their implications. By their nature, the simulation model assumptions and simplifications (3.4.2.7 of Chapter 3) can be viewed as limitations because they in one way or another restrict the accuracy of the model. Although validation of the baseline simulation model in this study was focused on demonstrating sufficient accuracy for the objectives at hand, this does not mean that it was without limitations. Other limitations are listed below:

i) The small proportion of valid incident location data available for modelling, as described in 3.6.2.1. Although this approach yielded spatial distributions of incidents that appeared to be approximately correct, a greater proportion of valid system data would make this aspect of the baseline model more accurate.
ii) The two-tier models used in this study consisted of PRVs and a mix of ALS and non-ALS ambulances. The data arising from these models thus does not answer questions about ALS provision at incident scenes when PRVs are the only ALS ESVs in the system, a scenario that exists in some urban EMS systems in South Africa. In such systems the provision of ALS at incidents would be expected to be much lower than that identified in the current study, assuming other aspects of the model are similar.

iii) The effect of increasing ESV numbers on other models (i.e. other than the dynamic single-tier model) was not assessed. Although this may not truly be a limitation (as it was not identified as an objective of the study), the decision to only investigate the effect of increasing ESV numbers on the best-performing model meant that information on this effect in all of the factor models was not available for comparison.

iv) The effect of human factors on efficiency of the dispatch process was not modelled. Several factors may have influenced the efficiency of dispatch decisions in the modelled system including variations in incident occurrence, fatigue and the additional dispatch workload of inter-hospital transfers which were not included in the baseline simulation model. Application of the policy to dispatch the closest ESV to an incident may have been consistently carried out by the simulation software, however human dispatchers would not have been able to match this level of efficiency or reliability. In some cases a conscious decision may have been made on the part of a dispatcher to deviate from a particular dispatch policy and these decisions were also not modelled.

5.8. FUTURE UTILITY AND VALUE OF THE BASELINE SIMULATION MODEL

The results and discussion in this study have been devoted to output data from the simulation models described in Chapter 3, which speak to objectives 1.6.2 (iii) and 1.6.3 of Chapter 1. However creation of the baseline simulation model was a primary objective of the study. Although the baseline model was intended to function as a departure point for the four factor models, and in turn for the generation of output data to be analysed, additional value of this model lies in the fact that it has some use beyond the scope of the current study.

Because the baseline simulation model is an approximation of a real system, data obtained from it cannot be generalised to any other real systems. The unique combination of dispatch and response processes, and the spatial relationships between ESV locations, incidents and hospitals are specific to EMS operations in Cape Town. Although this appears to be a severe restriction it is no different to the situation with regard to data and new knowledge arising from research on any other EMS simulation model, all of which are based on unique systems. The baseline simulation model can therefore still be used for a wide range of future research which may involve improvement, change or extension of this model in order to test a variety of hypotheses including more detailed modelling of dispatch centre processes.

The baseline simulation model also lends itself, albeit with significantly more development effort, to application in other systems. The simulation software used separates spatial layout of objects and process logic in such a way that it is not difficult to retain either one while making significant changes to the other. Specifically, changing the spatial relationships between ESVs, incidents and hospitals could be achieved fairly easily to reflect those of a different system. Likewise, process logic and input configurations could be modified to represent different dispatch and response policies. The basic framework represented in the baseline simulation model represents a reasonable template on top of which these changes could be made in future research.

For the reasons given above, the existing baseline simulation model has value as a future research tool in the area of ESM simulation research. As such it makes a unique contribution quite apart from being the source of output data analysed in this study.

5.9. SUMMARY

This Chapter presented the main study findings, each related to a part of the original research question. Theoretical and some financial implications of these findings were given, which centre mainly on the relationships between ESV numbers, availability, hand-off delays and response performance with references to the theory of constraints and queuing theory in order explain some of the findings and emphasise the need for further research.

Implications of the study findings for the design of EMS response systems in South Africa were discussed under eight headings ranging from the importance of decentralisation of ESVs and the futility of increasing ESV numbers to improve response performance, through to the need for more rigorous study of response system constraints and critique of the current national response benchmark values and their validity. These implications lead into the final Chapter of this thesis which gives a brief conclusion, followed by recommendations and a some ideas for further research in this area of inquiry.

CHAPTER 6: CONCLUSION

The time taken for EMS to respond to an incident requiring EMC is internationally acknowledged as an important system performance indicator, notwithstanding the relevance of other performance indicators related to patient care. There has been a long-standing inability of EMS systems in large urban centres of South Africa to meet response performance benchmarks despite interventions aimed at achieving this, the most common of which is to increase the size of ESV fleets. The importance of this study lies as much in what it shows is not possible, as in what it suggests should be done to improve response performance in a large urban South African EMS system. The dynamics of such systems are undeniably complex, as will be the strategies capable of improving their performance. It is up to leaders and decision-makers in South African EMS to look more critically at their systems, see the opportunity for real change and invest in solutions capable of making a real difference.

Several recommendations are now made, arising from sections 5.6.1 to 5.6.7 of the previous Chapter. These recommendations are followed by a brief discussion focused on possible areas of future research, in order to further investigate some of the questions identified in this study.

6.1. RECOMMENDATIONS

Recommendation 1

Positioning of ESVs relative to incident locations appears to have the greatest effect on response performance, based on results of the current study and those previously reported. EMS decision-makers wishing to significantly improve response performance within their systems should direct future research, development and planning efforts towards better understanding these effects and implementing change in this area. Although seen traditionally as a complex, and potentially costly, area of development because of the degree of expertise required, investment in this area of development is likely to produce the greatest long-term return.

Recommendation 2

In conjunction with Recommendation 1, improvement of response performance is likely to benefit from analysis of historical demand patterns in order to optimise ESV redeployment policies and to match ESV location and numbers with changing patterns of demand. In order to achieve this, it is a fundamental requirement of any large urban EMS system to have access to reliable historical incident location data, obtained from ESV tracking technology rather than geocoded incident

address data. These data should be used, together with a suitable analytical method, to devise an ESV redeployment policy aimed at minimising response distance according to predicted changes in demand. Computer simulation is an ideal method for analysis and development of such an approach, however this can only be done in the presence of reliable incident location data as referred to above.

Recommendation 3

Having emphasised the importance of response distance and decentralisation in response performance, the use of PRVs as they were modelled in the current study appears counterintuitive. PRVs cover larger areas and distances when responding and thus represent the opposite of the advantages associated with decentralisation. Although further research is needed in this area it would appear that PRVs, as deployed in the current study, add no value in enhancing response performance and should not be used in this way.

Recommendation 4

As recommended above, emphasis should be placed in large urban EMS systems on analysing and developing ESV location-related processes rather than assuming that adding ESVs to existing systems structure will produce better response performance. Each EMS system should strive to determine, on the basis of scientific data, what the optimal ESV resourcing levels are for their unique geo-spatial and systems design characteristics.

Recommendation 5

Unless all of the constraints in EMS systems are identified, it will not be possible to understand them and to re-align resource flow in order to exploit them. Instead of the traditional “trial-and-error” approach, improvement of response performance in EMS systems should be undertaken in a more scientifically rigorous way.

Recommendation 6

Analysis of dispatch centre processes and steps to minimise dispatch delays should be implemented before other parts of the system are changed. This should include all dispatch centre processes, including those devoted to inter-hospital transfers.

Recommendation 7

The national response time benchmarks require revision in line with what can be realistically expected in EMS systems of the type found in urban centres in South Africa, and wherever possible, in line with evidence of an effect on outcome. It seems pointless to set benchmarks that have never been achieved and, more importantly, are never likely to be. Alternatively, if the benchmarks are to remain in their current form, resources must be made available for EMS decision-makers to analyse their existing systems and to design and implement interventions capable of delivering this level of performance.

6.2. FUTURE RESEARCH

Much opportunity remains to further investigate a range of response system factors and how these influence response performance using simulation. Indeed, the existence of a validated simulation model of a South African urban EMS system makes such investigation more accessible than ever. Of the vast number of possible future studies, the following (phrased as questions) are of particular relevance:

- *What are the most important constraints on response performance in different response and ESV location models?*
- *How can existing incident demand forecasting methods be used in conjunction with simulation to assess their effectiveness?*
- *What is the most effective way of deploying and re-deploying ESVs in anticipation of demand in order to ensure a given response time at a given confidence level?*
- *What is the most efficient and cost-effective combination of response model, ESV location and ESV numbers that satisfies given response performance criteria?*
- *What is the value of PRVs in the provision of rapid access to ALS-level care for P1s?*
- *What are the cost implications of single- and two-tier response models?*
- *What is the effect of decreasing mission times for P1 and P2 cases on response performance?*

All of these questions could be asked within the context of a single type of simulation model, or across a range of models, as in the current study. Future simulation research on response performance could represent the interaction between dispatch processes, the rest of the EMS system and processes in the Emergency Centre by modelling dispatch and Emergency Centre processes in greater detail. This interaction in itself, especially that between capacity of Emergency Centres to accept patients and behaviour of the response system, is worthy of future study.

REFERENCES

1. Banks, J; Carson, JS; Nelson BND. Introduction to Discrete-Event Simulation. Discrete-Event System Simulation. Upper Saddle River: Prentice-Hall; 2010. p. 4–22.

2. Robinson S. Simulation Studies: An Overview. Simulation: The Practice of Model Development and Use. Chichester, West Sussex, England: John Wiley & Sons; 2004. p. 52–62.

3. Robinson S. Conceptual Modelling. Simulation: The Practice of Model Development and Use. Chichester, West Sussex, England: John Wiley & Sons; 2004. p. 63–75.

4. Robinson S. Simulation: What, Why and When? Simulation: The Practice of Model Development and Use. Chichester, West Sussex, England: John Wiley & Sons; 2004. p. 1–12.

5. Law A, Kelton DW. Basic Simulation Modeling. Simulation Modeling and Analysis. Third. Singapore: McGraw-Hill; 2000. p. 1–105.

6. Holliman J. Standard EMS Terms and Definitions. In: Tintinalli J, Cameron P, Holliman J, editors. EMS: A Practical Global Guidebook. Shelton, Connecticut: People's Medical Publishing House; 2010. p. 3–7.

7. Walz B, Krumperman K, Zigmont J. Glossary. Foundations of EMS Systems. New York: Delmar, Cengage Learning; 2011. p. 283–97.

8. Law A, Kelton DW. Building Valid, Credible, and Appropriately Detailed Simulation Models. Simulation Modeling and Analysis. Third. Singapore: McGraw-Hill; 2000. p. 264–91.

9. Banks, J; Carson, JS; Nelson BND. Verification, Calibration, and Validation of Simulation Models. Discrete-Event System Simulation. Upper Saddle River: Prentice-Hall; 2010. p. 388–416.

10. Law A, Kelton DW. Simulation Software. Simulation Modeling and Analysis. Third. Singapore: McGraw-Hill; 2000. p. 202–34.

11. Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington D.C.; 2001.

12. Delbridge TR. EMS Systems. In: Bass R, Brice J, Delbridge TR, Gunderson M, editors. Medical oversight of EMS. Dubuque, Iowa: Kendall Hunt Professional; 2009. p. 151–79.

13. Walz B, Krumperman K, Zigmont J. Introduction to Emergency Medical Systems. Foundations of EMS Systems. New York: Delmar, Cengage Learning; 2011. p. 2–12.

14. National Highway Traffic Safety Adminstration: EMS. What is EMS? [Internet]. [cited 2013 Nov 21]. Available from: http://www.ems.gov/whatisEMS.htm

15. Meislin HW, Conn JB, Conroy C, Tibbitts M. Emergency Medical Service Agency Definitions of Response Itervals. Annals of Emergency Medicine. 1999 Oct;34(4 Pt 1):453–8.

16. Spaite DW, Valenzuela TD, Meislin HW, Criss EA, Hinsberg P. Prospective Validation of a New Model for Evaluating Emergency Medical Services Systems by In-field Observation of Specific Time Intervals in Prehospital Care. Annals of Emergency Medicine. 1993;22(4):638–45.

17. Eisenberg MS, Bergner L, Hallstrom A. Cardiac Resuscitation in the Community. Importance of Rapid Provision and Implications for Program Planning. Journal of the American Medical Association. 1979 May 4;241(18):1905–7.

18. Mullie A, Van Hoeyweghen R, Quets A. Influence of Time Intervals on Outcome of CPR. The Cerebral Resuscitation Study Group. Resuscitation. 1989 Jan;17 Suppl:S23–33; discussion S199–206.

19. Sampalis JS, Lavoie A, Williams JI, Mulder DS, Kalina M. Impact of On-site Care, Prehospital Time, and Level of In-hospital Care on Survival in Severely Injured Patients. The Journal of Trauma. 1993 Feb;34(2):252–61.

20. Dunford J, Domeier RM, Blackwell T, Mears G, Overton J, Rivera-Rivera EJ, et al. Performance Measurements in Emergency Medical Services. Prehospital Emergency Care. 2002;6(1):92–8.

21. Price L. Treating the Clock and Not the Patient: Ambulance Response Times and Risk. Quality & Safety in Health Care. 2006 Apr;15(2):127–30.

22. Pons P, Markovchick VJ. Eight Minutes or Less: Does the Ambulance Response Time Guideline Impact Trauma Patient Outcome? The Journal of Emergency Medicine. 2002;23(1):43–8.

23. National Fire Protection Association. NFPA 1710 Standard for the Organization and Deployment of Fire Suppression Operations , Emergency Medical Operations , and Special Operations to the Public by Career Fire Departments 2001 Edition. National Fire Protection Association; 2001. p. 1710–9.

24. Blackwell TH, Kaufman JS. Response Time Effectiveness: Comparison of Response Time and Survival in an Urban Emergency Medical Services System. Academic Emergency Medicine. 2002 Apr;9(4):288–95.

25. Blackwell TH, Kline J a, Willis JJ, Hicks GM. Lack of Association Between Prehospital Response Times and Patient Outcomes. Prehospital Emergency Care. 2009;13(4):444–50.

26. Pons PT, Haukoos JS, Bludworth W, Cribley T, Pons K a, Markovchick VJ. Paramedic Response Time: Does it Affect Patient Survival? Academic Emergency Medicine. 2005 Jul;12(7):594–600.

27. Swor R a, Cone DC. Emergency Medical Services Advanced Life Support Response Times: Lots of Heat, Little Light. Academic Emergency Medicine. 2002 Apr;9(4):320–1.

28. Weiss S, Fullerton L, Oglesbee S, Duerden B, Froman P. Does Ambulance Response Time Influence Patient Condition Among Patients With Specific Medical and Trauma Emergencies? Southern Medical Journal. 2013 Mar;106(3):230–5.

29. Stout J, Pepe PE, Mosesso VN. All-Advanced Life Support vs Tiered-Response Ambulance Systems. Prehospital Emergency Care. 2000;4(1):1–6.

30. Stout J. System Status Management: The Fact Is, Its Everywhere [Internet]. 1989 [cited 2012 Apr 13]. Available from: http://www.stouts.org/jack/EMS/JEMS0489.htm

31. Western Cape Department of Health. Western Cape Department of Health Annual Report 2012-2013 [Internet]. 2013. Available from: http://www.westerncape.gov.za/dept/health/documents/annual_reports/2012

32. Gauteng Department of Health and Social Development. Gauteng Department of Health and Social Development Annual Report 2012-2013 [Internet]. 2013. Available from: http://www.health.gpg.gov.za/Document/Pages/AnnualReports.aspx

33. KwaZulu-Natal Department of Health. KwaZulu-Natal Department of Health Annual Report 2012-2013; Part B: Perfromance Information [Internet]. 2013. Available from: http://www.kznhealth.gov.za/2012.2013_annual_report.htm

34. Funder KS, Petersen JA, Steinmetz J. On-scene Time and Outcome After Penetrating Trauma: An Observational Study. Emergency Medicine Journal. 2011 Sep;28(9):797–801.

35. Gonzalez RP, Cummings GR, Phelan HA, Mulekar MS, Rodning CB. Does Increased Emergency Medical Services Prehospital Time Affect Patient Mortality in Rural Motor Vehicle Crashes? A Statewide Analysis. American Journal of Surgery. 2009 Jan;197(1):30–4.

36. Yasunaga H, Miyata H, Horiguchi H, Tanabe S, Akahane M, Ogawa T, et al. Population Density, Call-response Interval, and Survival of Out-of-hospital Cardiac Arrest. International Journal of Health Geographics. 2011 Jan;10:26.

37. Wainer G. Modeling and Simulation Concepts. Discrete-event Modeling and Simulation: A Practitioner's Approach. Boca Raton: CRC Press; 2009.

38. Kallsen G, Stroh G. Quality in Perspective. In: Lerner EB, Pirallo RG, Swor R, White L, editors. Evaluating and Improving Quality in EMS. Dubuque, Iowa: Kendall Hunt Professional; 2009. p. 3–11.

39. Chassin MR, Galvin RW. The Urgent Need to Improve Health Care Quality. Institute of Medicine National Roundtable on Health Care Quality. Journal of the American Medical Association. 1998 Sep 16;280(11):1000–5.

40. Moore L. Performance Measurement in EMS. In: Lerner E, Pirallo R, Swor R, White L, editors. Evaluating and Improving Quality in EMS. Dubuque, Iowa: Kendall Hunt Professional; 2009. p. 80–98.

41. Smith K, Barger B, Currell A. Development of a an EMS Quality Improvement Program. In: Tintinalli, J;Cameron, P;Holliman C, editor. EMS: A Practical Global Guidebook. Shelton, Connecticut: People's Medical Publishing House; 2010. p. 196–218.

42. Hosken G. Pretoria's Ambulance Service "in shambles". [Internet]. Independent Online. 2007 [cited 2012 Apr 14]. Available from: http://www.iol.co.za/news/south-africa/pretoria-s-ambulance-service-in-shambles-1.371317

43. Bloom J. Joburg Emergency Response Times Worse [Internet]. Democratic Alliance Gauteng. 2012 [cited 2012 Apr 14]. Available from: http://dagauteng.wordpress.com/tag/response-times/

44. Du Plessis C. Critical Response Times Slow Down [Internet]. Times Live. 2011 [cited 2012 Apr 14]. Available from: http://www.timeslive.co.za/local/2011/11/07/critical-response-times-slow-down

45. Myers B. Tiered EMS Systems. In: Bass R, Brice J, Delbridge T, Gunderson M, editors. Medical Oversight of EMS. Dubuque, Iowa: Kendall Hunt Professional; 2009. p. 180–6.

46. Morse A. Transforming NHS Ambulance Services. Norwich; 2011.

47. Krafft T, García Castrillo-Riesgo L, Edwards S, Fischer M, Overton J, Robertson-Steel I, et al. European Emergency Data Project (EED Project): EMS Data-based Health Surveillance System. European Journal of Public Health. 2003 Sep;13(3 Suppl):85–90.

48. Overton J. Benchmarking EMS in Europe. 7th Congress of the European Resuscitation Council [Internet]. 2004. Available from: http://www.eed-network.eu/assets/presentations/EED_Benchmarking.pdf

49. Steering Committee for the Review of Government Service provision. Report on Government Services. Melbourne; 2012.

50. Meara PO. A Generic Performance Framework for Ambulance Services : An Australian Health Services Perspective. Journal of Emergency Primary Health Care. 2012;3(3):1–13.

51. Al-Shaqsi S. Response Time as a Sole Performance Indicator in EMS : Pitfalls and Solutions. Open Access Emergency Medicine. 2010;2:1–6.

52. Greenberg MD, Garrison HG, Delbridge TR, Miller WR, Mosesso VN, Roth RN, et al. Quality Indicators for Out-of-hospital Emergency Medical Services: The Paramedics' Perspective. Prehospital Emergency Care. 1997;1(1):23–7.

53. Myers JB, Slovis CM, Eckstein M, Goodloe JM, Isaacs SM, Loflin JR, et al. Evidence-based Performance Measures for Emergency Medical Services Systems: A Model for Expanded EMS Benchmarking. Prehospital Emergency Care. 2007;12(2):141–51.

54. Wankhade P. Performance Measurement and the UK Emergency Ambulance Service: Unintended Consequences of the Ambulance Response Time Targets. International Journal of Public Sector Management. 2011;24(5):384–402.

55. Wilde ET. Do Emergency Medical System Response Times Matter for Health Outcomes? Health Economics. 2013;806(June 2012):790–806.

56. Clark DE, Winchell RJ, Betensky R a. Estimating the Effect of Emergency Care on Early Survival After Traffic Crashes. Accident Analysis and Prevention. Elsevier Ltd; 2013 Nov;60:141–7.

57. Sánchez-Mangas R, García-Ferrrer A, de Juan A, Arroyo AM. The Probability of Death in Road Traffic Accidents. How Important is a Quick Medical Response? Accident Analysis and Prevention. 2010 Jul;42(4):1048–56.

58. Blanchard IE, Doig CJ, Hagel BE, Anton AR, Zygun DA, Kortbeek JB, et al. Emergency Medical Services Response Time and Mortality in an Urban Setting. Prehospital Emergency Care. 2011;16(1):142–51.

59. Banks, J; Carson, JS; Nelson BND. General Principles. Discrete-Event System Simulation. Upper Saddle River: Prentice-Hall; 2010. p. 88–116.

60. Robinson S. Inside Simulation Software. Simulation: The Practice of Model Development and Use. Chichester, West Sussex, England: John Wiley & Sons; 2004. p. 13–36.

61. Banks, J; Carson, JS; Nelson BND. Random-variate Generation. Discrete-Event System Simulation. Upper Saddle River: Prentice-Hall; 2010. p. 299–332.

62. Banks, J; Carson, JS; Nelson BND. Input Modeling. Discrete-Event System Simulation. Upper Saddle River: Prentice-Hall; 2010. p. 335–87.

63. Kelton D, Smith J, Sturrock D. Input Analysis. Simio & Simulation: Modeling, Analysis and Applications. Second. USA: Learning Solutions; 2011. p. 67–95.

64. Law A, Kelton DW. Random-number Generators. Simulation Modeling and Analysis. Third. Singapore: McGraw-Hill; 2000. p. 402–36.

65. Matsumoto M, Nishimura T. Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-random Number Generator. ACM Transactions on Modeling and Computer Simulation. 1998;8:3–30.

66. Robinson S. Model Coding. Simulation: The Practice of Model Development and Use. Chichester, West Sussex, England: John Wiley & Sons; 2004. p. 127–36.

67. Banks, J; Carson, JS; Nelson BND. Estimation of Relative Performance. Discrete-Event System Simulation. Upper Saddle River: Prentice-Hall; 2010. p. 463–506.

68. Law A, Kelton DW. Variance Reduction Techniques. Simulation Modeling and Analysis. Third. Singapore: McGraw-Hill; 2000. p. 581–621.

69. Robinson S. Conceptual Modeling for Simulation. Conceptual Modeling for Discrete-Event Simulation. CRC Press; 2010. p. 3–30.

70. Robinson S. Simulation: The practice of model development and use. Chichester, West Sussex, England: John Wiley & Sons; 2004.

71. Robinson S. A Framework for Simulation Conceptual Modeling. Conceptual Modeling for Discrete-Event Simulation [Internet]. CRC Press; 2010. p. 73–101. Available from: http://dx.doi.org/10.1201/9781439810385-c4

72. Robinson S. Data Collection and Analysis. Simulation: The Practice of Model Development and Use. Chichester, West Sussex, England: John Wiley & Sons; 2004. p. 95–125.

73. Kuhl ME, Ivy JS, Lada EK, Steiger NM, Wagner MA, Wilson JR. Univariate Input Models for Stochastic Simulation. Journal of Simulation. Palgrave Macmillan; 2010 Feb 26;4(2):81–97.

74. Simio LLC. Pert (Distributions). Simio LLC; 2013.

75. Robinson S. Verification, Validation and Confidence. Simulation: The Practice of Model Development and Use. Chichester, West Sussex, England: John Wiley & Sons; 2004. p. 209–24.

76. Sargent RG. Verification and Validation of Simulation Models. Proceedings of the 2011 Winter Simulation Conference S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, eds. 2011. p. 183–98.

77. Law A, Kelton DW. Output Data Analysis for a Single System. Simulation Modeling and Analysis. Third. Singapore: McGraw-Hill; 2000. p. 496–552.

78. Robinson S. Experimentation: Obtaining Accurate Results. Simulation: The Practice of Model Development and Use. Chichester, West Sussex, England: John Wiley & Sons; 2004. p. 136–68.

79. Banks, J; Carson, JS; Nelson BND. Estimation of Absolute Performance. Discrete-Event System Simulation. Upper Saddle River: Prentice-Hall; 2010. p. 417–62.

80. Robinson S. Experimentation: Searching the Solution Space. Simulation: The Practice of Model Development and Use. Chichester, West Sussex, England: John Wiley & Sons; 2004. p. 168–99.

81. Law A, Kelton DW. Comparing Alternative System Configurations. Simulation Modeling and Analysis. Singapore: McGraw-Hill; 2000. p. 553–80.

82. Aboueljinane L, Sahin E, Jemai Z. A Review on Simulation Models Applied to Emergency Medical Service Operations. Computers & Industrial Engineering. Elsevier Ltd; 2013 Dec;66(4):734–50.

83. Sacco WJ, Navin DM, Fiedler KE, Waddell RK, Long WB, Buckman RF. Precise Formulation and Evidence-based Application of Resource-constrained Triage. Academic Emergency Medicine. 2005 Aug;12(8):759–70.

84. Wang Y, Luangkesorn KL, Shuman L. Modeling Emergency Medical Response to a Mass Casualty Incident Using Agent Based Simulation. Socio-Economic Planning Sciences. 2012 Dec;46(4):281–90.

85. Inoue H, Yanagisawa S, Kamae I. Computer-simulated Assessment of Methods of Transporting Severely Injured Individuals in Disaster—Case Study of an Airport Accident. Computer Methods and Programs in Biomedicine. 2006 Mar;81(3):256–65.

86. Harewood SI. Emergency Ambulance Deployment in Barbados: A Multi-objective Approach. Journal of the Operational Research Society. 2002 Feb;53(2):185–92.

87. Silva PM, Pinto LM. Emergency Medical Systems Analysis by Simulation and Optimisation. Proceedings of the 2010 Winter Simulation Conference B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds. 2010. p. 2422–32.

88. Van Buuren M, Aardal K, van der Mei R, Post H. Evaluating the Dynamic Dispatch Strategies for Emergency Medical Services: TIFAR Simulation Tool. Proceedings of the 2012 Winter Simulation Conference C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, eds. 2012. p. 509–19.

89. Fitzsimmons J. An Emergency Medical System Simulation Model. Proceedings of the 1971 Winter Simulation Conference. 1971. p. 19–25.

90. Fitzsimmons JA. A Methodology for Emergency Ambulance Deployment. Management Science. 1973;19(6):627–36.

91. Swoveland C, Uyeno D, Vertinsky I, Vickson R. A Simulation-based Methodology for Optimization of Ambulance Service Policies. Socio-Economic Planning Sciences. 1973 Dec;7(6):697–703.

92. Lubicz M, Mielczarek B. Simulation Modelling of Emergency Medical Services. European Journal of Operational Research. 1987 May;29(2):178–85.

93. Repede JF, Bernardo JJ. Developing and Validating a Decision Support System for Locating Emergency Medical Vehicles in Louisville, Kentucky. European Journal of Operational Research. 1994 Jun 30;75(3):567–81.

94. Ingolfsson A, Erkut E, Budge S. Simulation of Single Start Station for Edmonton EMS. Journal of the Operational Research Society. 2003 Jul;54(7):736–46.

95. Maxwell MS, Henderson SG. Ambulance Redeployment: An Approximate Dynamic Programming Approach. Proceedings of the 2009 Winter Simulation Conference M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds. 2009. p. 1850–60.

96. Aboueljinane L. Reducing Ambulance Response Time Using Simulation: The Case of Val- De-Marne Department Emergency Medical Service. Proceedings of the 2012 Winter Simulation Conference C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher, eds. 2012. p. 943–54.

97. Berlin GN, Liebman JC. Mathematical Analysis of Emergency Ambulance Location. Socio-Economic Planning Sciences. 1974 Dec;8(6):323–8.

98. Goldberg J, Dietrich R, Chen JM, Mitwasi M, Valenzuela T, Criss E. A simulation Model for Evaluating a Set of Emergency Vehicle Base Locations: Development, Validation, and Usage. Socio-economic Planning Sciences. 1990 Jan;24(2):125–41.

99. Aringhieri R, Carello G, Morale D. Ambulance Location Through Optimisation and Simulation: The Case of Milano Urban Area. The 38th Annual Conference of the Italian Operations Research Society Optimization and Decision Sciences. 2007.

100. Henderson SG, Mason AJ. Estimating Ambulance Requirements in Auckland, New Zealand. Proceedings of the 1999 Winter Simulation Conference P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds. 1999. p. 1670–4.

101. Wei Lam SS, Zhang ZC, Oh HC, Ng YY, Wah W, Hock Ong ME. Reducing Ambulance Response Times Using Discrete Event Simulation. Prehospital Emergency Care. 2014;18(2):207–16.

102. Inakawa K. Effect of Ambulance Station Locations and Number of Ambulances to the Quality of the Emergency Service. The Ninth International Symposium on Operations Research and Its Applications (ISORA'10). 2010. p. 340–7.

103. Uyeno DH, Seeberg C. A Practical Methodology for Ambulance Location. Simulation. 1984 Aug 1;43(2):79–87.

104. Trudeau P, Rosseau J-M, Ferland J, Choquette J. An Operations Research Approach for the Planning and Operation of an Ambulance Service. Information Systems and Operational Research. 1989;27(1):95–113.

105. Wears R, Winton C. Simulation Modeling of Prehospital Trauma Care. Proceedings of the 1993 Winter Simulation Conference. 1993. p. 1216–24.

106. Savas E. Simulation and Cost-effectiveness Analysis of New York's Emergency Ambulance Service. Management Science. 1968;18(12):B608–B627.

107. Su S, Shih C-L. Resource Reallocation in an Emergency Medical Service System Using Computer Simulation. The American Journal of Emergency Medicine. 2002 Nov;20(7):627–34.

108. Gunes E, Szechtman R. A Simulation Model of a Helicopter Ambulance Service. Proceedings of the 2005 Winter Simulation Conference M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds. 2005. p. 951–7.

109. Peleg K, Pliskin J. A Geographic Information System Simulation Model of EMS: Reducing Ambulance Response Time. The American Journal of Emergency Medicine. 2004 May;22(3):164–70.

110. Iskander W. Simulation Modeling for Emergency Medical Service Systems. Proceedings of the 1989 Winter Simulation Conference. 1989. p. 1107–11.

111. Rajagopalan HK, Saydam C, Xiao J. A Multiperiod Set Covering Location Model for Dynamic Redeployment of Ambulances. Computers & Operations Research. 2008 Mar;35(3):814–26.

112. Wu C-H, Hwang KP. Using a Discrete-event Simulation to Balance Ambulance Availability and Demand in Static Deployment Systems. Academic Emergency Medicine. 2009 Dec;16(12):1359–66.

113. Koch O, Weigl H. Modeling the Ambulance Service of the Austrian Red Cross. Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.03EX693). Ieee; 2003. p. 1701–6.

114. Su S, Shih C-L. Modeling an Emergency Medical Services System Using Computer Simulation. International Journal of Medical Informatics. 2003 Dec;72(1-3):57–72.

115. Creswell J. Quantitative Methods. Resesearch design: Qualitative, quantitative and mixed methods approaches. Third. Thousand Oaks, California: Sage Publications; 2009. p. 145–71.

116. Montgomery D. Introduction. Design and Analysis of Experiments. Fifth. New Jersey: John Wiley & Sons; 2001. p. 1–19.

117. Pegden CD. Intelligent Objects : The Future of Simulation. Simio, LLC;

118. Kelton D, Smith J, Sturrock D. Working With Model Data. Simio & Simulation: Modeling, Analysis and Applications. Second. USA: Learning Solutions; 2011. p. 203–34.

119. Garson D. GLM Multivariate, MANOVA & MANCOVA. Asheboro, North Carolina: Statistical Associates Publishing; 2012. p. 40–7.

120. Rahman S. Theory of Constraints: A Review of the Philosophy and its Applications. International Journal of Operations and Production Management. 1998;18(4):336–55.

121. Institute of Management Accountants. Theory of Constraints: Management System Fundamentals. Montvale, New Jersey: Institute of Management Accountants; 1999.

122. Zilm F, Crane J, Roche K. New Directions in Emergency Service Operations and Planning. Journal of Ambulatory Care Management. 2010;33(4):296–306.

123. Banks, J; Carson, JS; Nelson BND. Queueing Models. Discrete-Event System Simulation. Upper Saddle River: Prentice-Hall; 2010. p. 227–74.

124. Law A, Kelton DW. Appendix 1B: A Primer on Queueing Systems. Simulation Modeling and Analysis. Singapore: McGraw-Hill; 2000. p. 94–6.

125. Kelton DW, Smith J, Sturrock D. Basics of Queueing Theory. Simio & Simulation: Modeling, Analysis and Applications. Second. USA: Learning Solutions; 2011. p. 21–35.

126. Brown WE, Dickison PD, Misselbeck WJA, Levine R. Longitudinal Emergency Medical Technician Atrribute and Demographic Study (LEADS): An Interim Report. Prehospital Emergency Care. 2002;6:433–9.

127. Chapman SA, Blau G, Pred R, Lopez AB. Correlates of Intent to Leave Job and Profession for Emergency Medical Technicians and Paramedics. Career Development International. 2009;14(5):487–503.

128. Hackland S, Stein C. Factors Influencing the Departure of South African Advanced Life Support Paramedics From Pre-hospital Operational Practice. African Journal of Emergency Medicine. 2011;1:62–8.

129. Kolesar P, Blum E. Square Root Laws for Fire Engine Response Distances. Management Science. 1973;19(12):1368–78.

130. Green L, Kolesar P. Improving Emergency Responsiveness with Management Science. Management Science. 2004;50(8):1001–14.

ANNEXURE A. Mean Incident Hourly Rates: Sector 1 (Groote Schuur Hospital)

Table A.1

Hour	0	1	2	3	4	5	6	7	8	9	10	11
Mon	2	1	1	1	2	1	2	2	2	2	2	2
Tue	1	1	1	1	1	1	2	2	2	2	2	2
Wed	1	1	1	2	1	1	2	2	3	2	2	2
Thu	1	1	1	1	1	1	1	2	2	2	2	2
Fri	1	1	1	1	1	1	2	2	3	2	2	2
Sat	1	2	1	1	1	1	1	2	2	2	2	2
Sun	2	2	2	1	1	1	1	2	2	2	2	3

Table A.2

Hour	12	13	14	15	16	17	18	19	20	21	22	23
Mon	2	2	2	2	2	2	2	3	2	2	2	2
Tue	2	2	2	2	2	2	2	2	2	2	2	1
Wed	2	2	2	2	2	2	2	2	2	2	1	2
Thu	2	2	2	2	2	2	2	2	2	2	2	1
Fri	2	2	2	2	2	2	2	2	2	2	2	2
Sat	2	2	2	2	2	2	2	3	2	2	2	2
Sun	2	2	2	3	2	2	2	3	2	2	2	2

ANNEXURE B. Mean Incident Hourly Rates: Sector 2 (GF Jooste Hospital)

Table B.1

Hour	0	1	2	3	4	5	6	7	8	9	10	11
Mon	6	6	4	4	3	2	3	4	5	5	5	6
Tue	4	3	3	3	2	2	3	5	6	6	6	6
Wed	3	3	2	2	2	2	3	4	5	5	6	5
Thu	3	3	2	2	1	2	3	4	6	5	5	5
Fri	3	3	2	3	2	2	3	3	4	5	5	6
Sat	3	3	2	3	2	2	2	4	4	5	5	5
Sun	5	4	3	3	3	2	2	3	3	4	5	5

Table B.2

Hour	12	13	14	15	16	17	18	19	20	21	22	23
Mon	5	5	4	5	5	5	7	8	7	7	6	5
Tue	6	5	4	4	4	4	5	7	6	6	6	4
Wed	5	4	4	4	4	3	4	6	5	6	5	4
Thu	4	4	4	4	4	4	5	5	5	6	5	4
Fri	4	4	4	4	4	4	5	5	5	5	5	4
Sat	4	4	5	4	4	4	4	5	6	6	6	5
Sun	5	5	5	4	4	5	6	7	7	8	8	8

ANNEXURE C. Mean Incident Hourly Rates: Sector 3 (Tygerberg Hospital)

Table C.1

Hour	0	1	2	3	4	5	6	7	8	9	10	11
Mon	8	9	7	6	5	4	4	6	7	8	8	8
Tue	5	5	3	3	3	3	5	7	8	9	9	9
Wed	4	4	4	3	3	3	4	7	7	9	8	8
Thu	3	4	3	3	2	3	3	6	7	8	7	7
Fri	4	3	3	3	3	3	3	7	7	7	8	8
Sat	4	4	3	3	3	3	4	6	6	7	8	8
Sun	6	7	5	5	4	3	4	5	6	7	7	8

Table C.2

Hour	12	13	14	15	16	17	18	19	20	21	22	23
Mon	8	7	8	7	8	8	10	11	10	11	9	8
Tue	8	7	7	6	6	7	8	10	8	9	7	5
Wed	6	6	5	6	5	7	7	8	8	8	7	6
Thu	7	5	6	6	6	6	7	8	6	8	7	5
Fri	6	7	6	6	6	7	7	9	7	8	6	6
Sat	7	6	6	6	6	7	7	9	8	9	9	8
Sun	7	7	7	7	8	8	10	12	11	13	13	11

ANNEXURE D. Mean Incident Hourly Rates: Sector 4 (Victoria Hospital)

Table D.1

Hour	0	1	2	3	4	5	6	7	8	9	10	11
Mon	3	3	2	2	2	2	2	3	2	4	4	3
Tue	2	2	2	2	1	2	2	3	3	3	4	3
Wed	1	2	1	2	1	1	2	3	3	4	3	4
Thu	2	2	2	2	1	1	2	3	2	3	3	3
Fri	2	2	1	2	1	2	2	3	3	3	3	3
Sat	2	2	1	1	1	1	2	2	2	3	3	3
Sun	2	2	2	2	2	1	2	2	3	3	3	3

Table D.2

Hour	12	13	14	15	16	17	18	19	20	21	22	23
Mon	4	3	3	3	3	3	4	4	4	4	3	2
Tue	3	3	3	2	3	3	3	4	3	3	3	2
Wed	3	3	3	2	3	3	3	3	3	3	3	2
Thu	3	2	3	2	3	3	3	3	3	3	2	2
Fri	3	3	2	2	3	3	3	3	3	3	2	2
Sat	3	3	2	3	2	3	3	4	3	3	3	3
Sun	3	3	3	4	3	3	4	5	4	4	4	4

ANNEXURE E. Turing Test Reports

E.1 Simulation Data

PRE-HOSPITAL TIMES REPORT

(Min)

Priority 1

	Mean	*95% CI*
Total Response Time	16.35	16.05; 16.65
Travel Response Time	10.52	10.35; 10.69
Scene Time	40.16	39.91; 40.41
Transport Time	21.92	21.71; 22.12
% Response <= 15 Minutes	65%	

Priority 2

	Mean	*95% CI*
Total Response Time	66.31	65.56; 67.05
Travel Response Time	16.61	16.34; 16.88
Scene Time	40.27	40.10; 40.44
Transport Time	20.60	20.41; 20.79
% Response <= 60 Minutes	84%	

E.2 System Data

PRE-HOSPITAL TIMES REPORT

(Min)

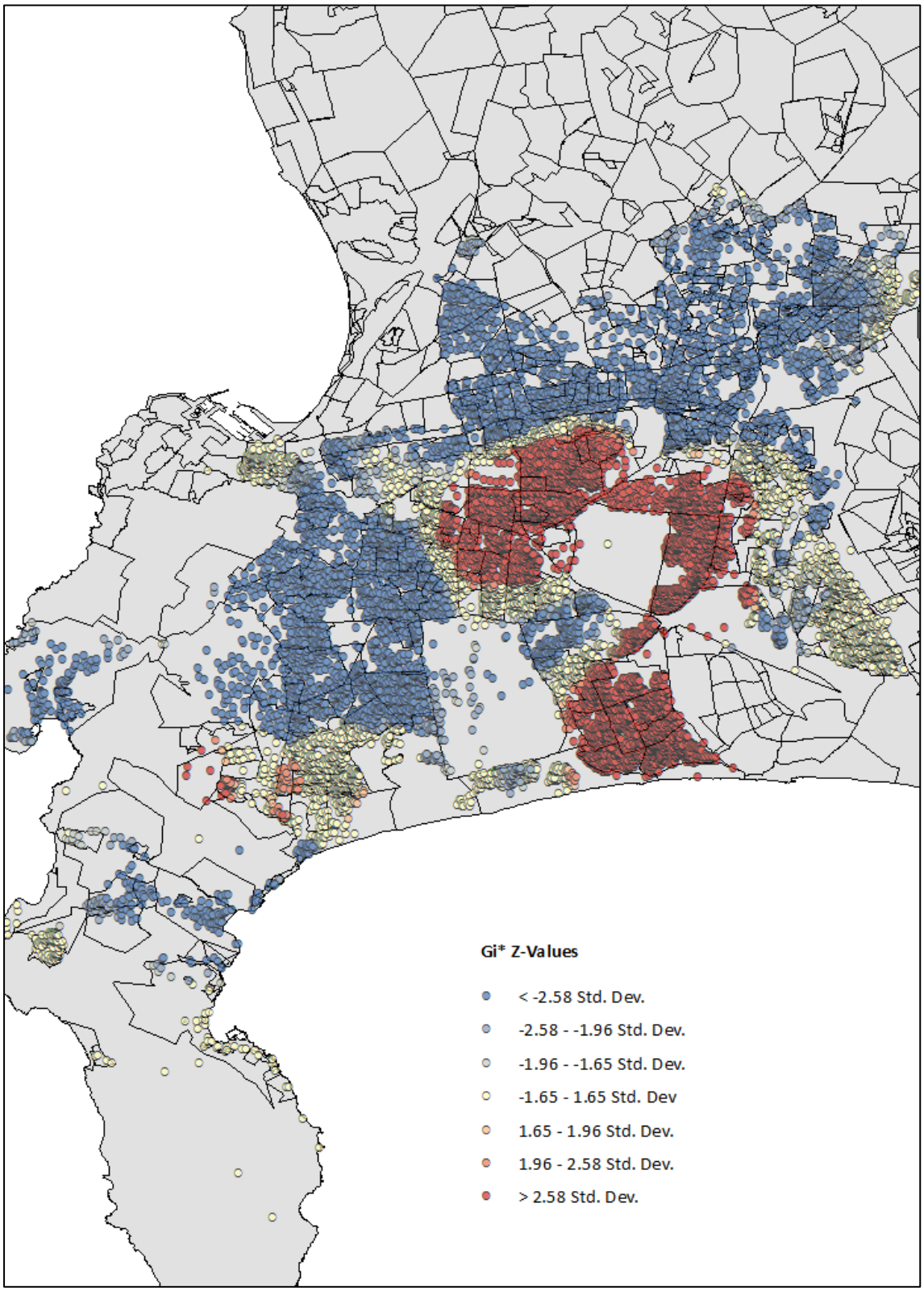
Priority 1

	Mean	*95% CI*
Total Response Time	16.44	16.19; 16.69
Travel Response Time	10.66	10.52; 10.81
Scene Time	40.19	39.59; 40.78
Transport Time	22.03	21.64; 22.42
% Response <= 15 Minutes	67%	

Priority 2

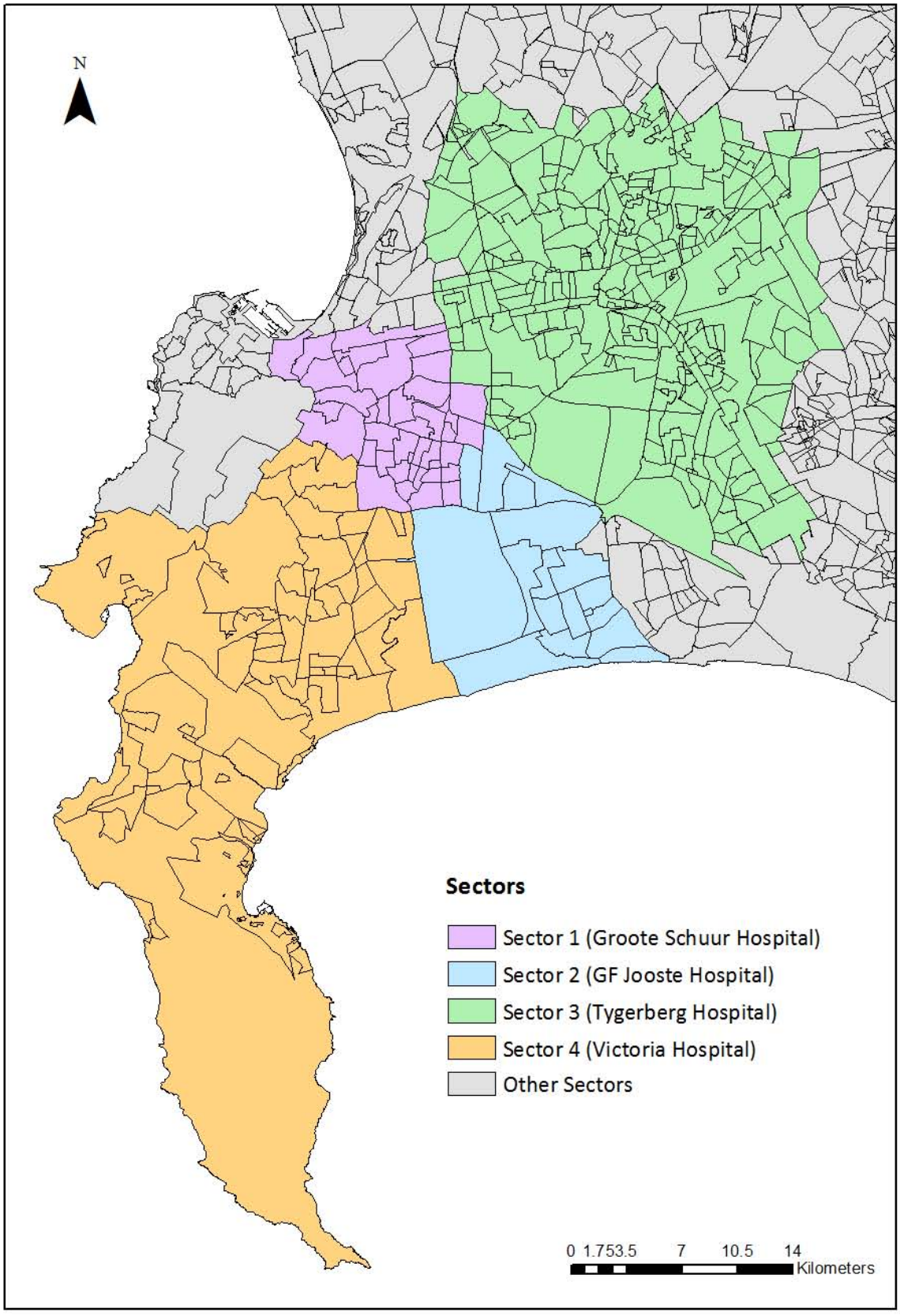
	Mean	*95% CI*
Total Response Time	67.17	66.10; 68.23
Travel Response Time	16.79	16.57; 17.01
Scene Time	40.20	39.68; 40.72
Transport Time	20.48	20.19; 20.78
% Response <= 60 Minutes	63%	

ANNEXURE F. Incident Locations: Hotspot Analysis

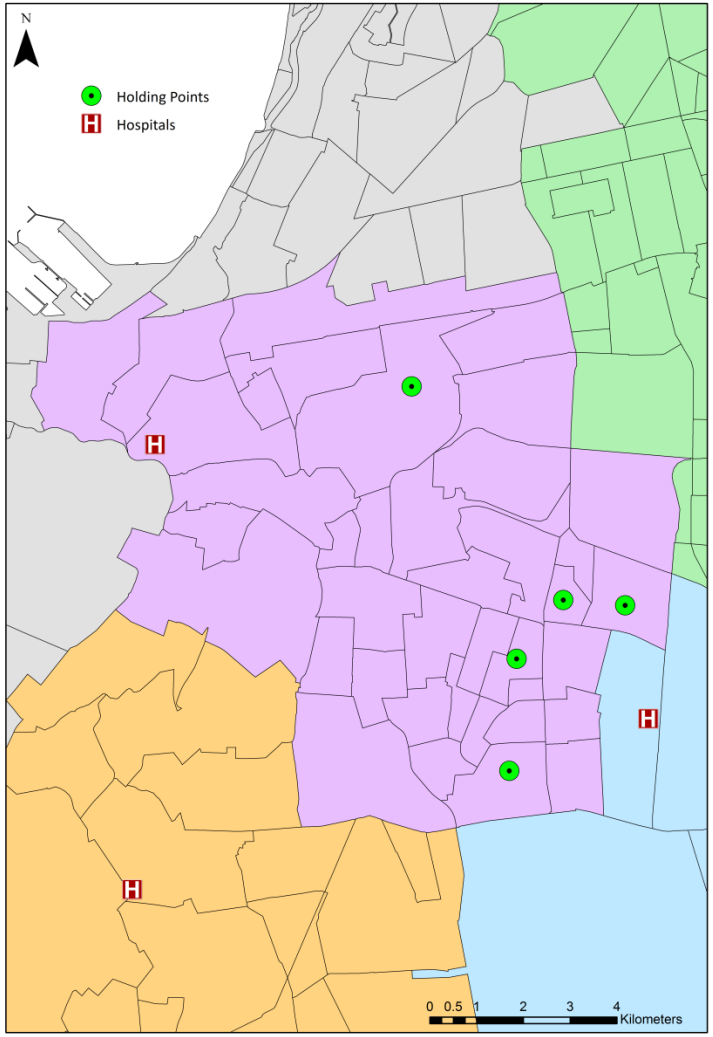


The Gi statistic is a Z statistic reflecting the intensity of clustering of points. Values greater than 1.96 or less than -1.96 are associated with p-values less than 0.05 suggesting that the intensity of clustering is not due to chance alone. Red areas are those where high numbers of incidents are surrounded by other areas with significantly high numbers of incidents and the opposite for blue. Clustering in yellow areas is not significant.

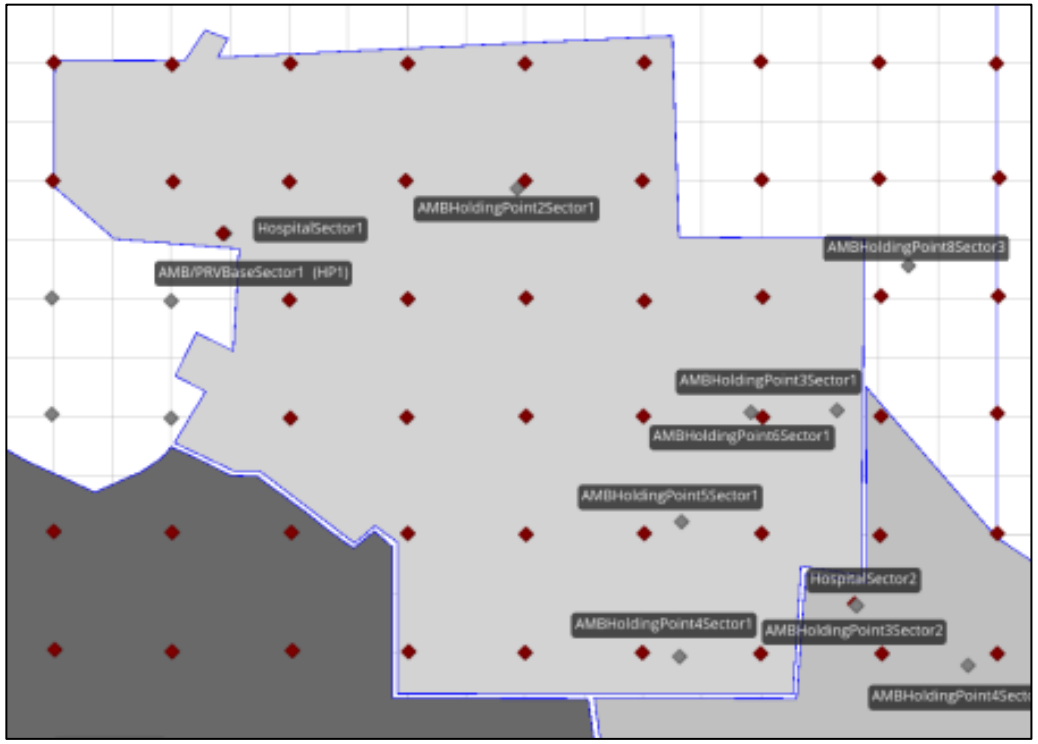
ANNEXURE G. All Sectors Included in the Simulation Model



ANNEXURE H. Sector 1 (Groote Schuur Hospital): Map and Modelled Sector

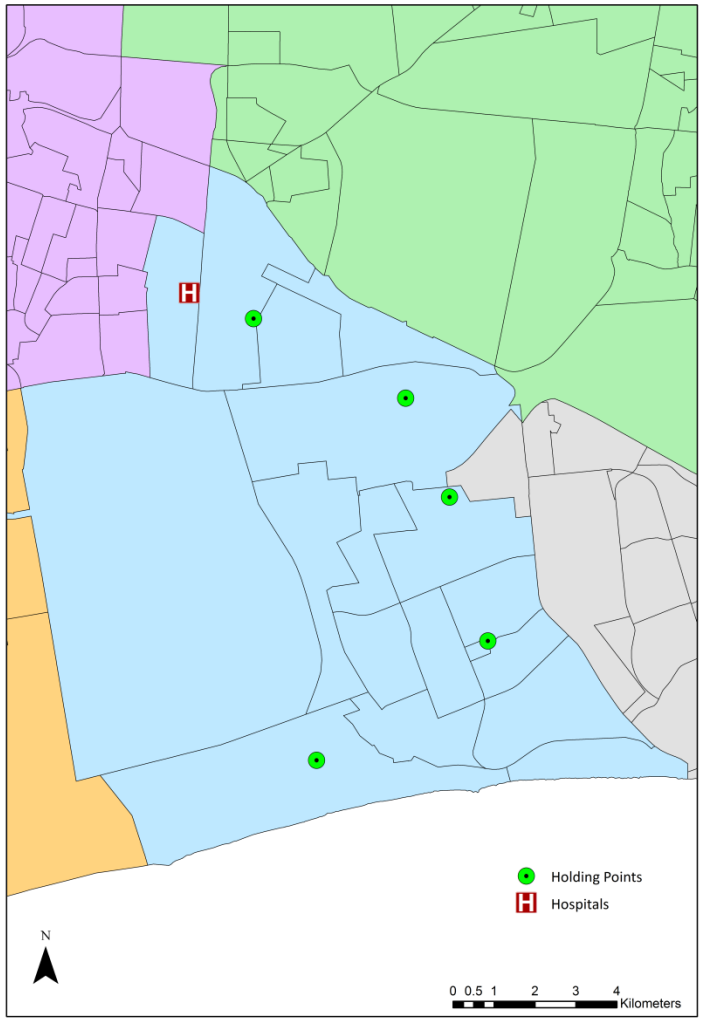


GIS Map

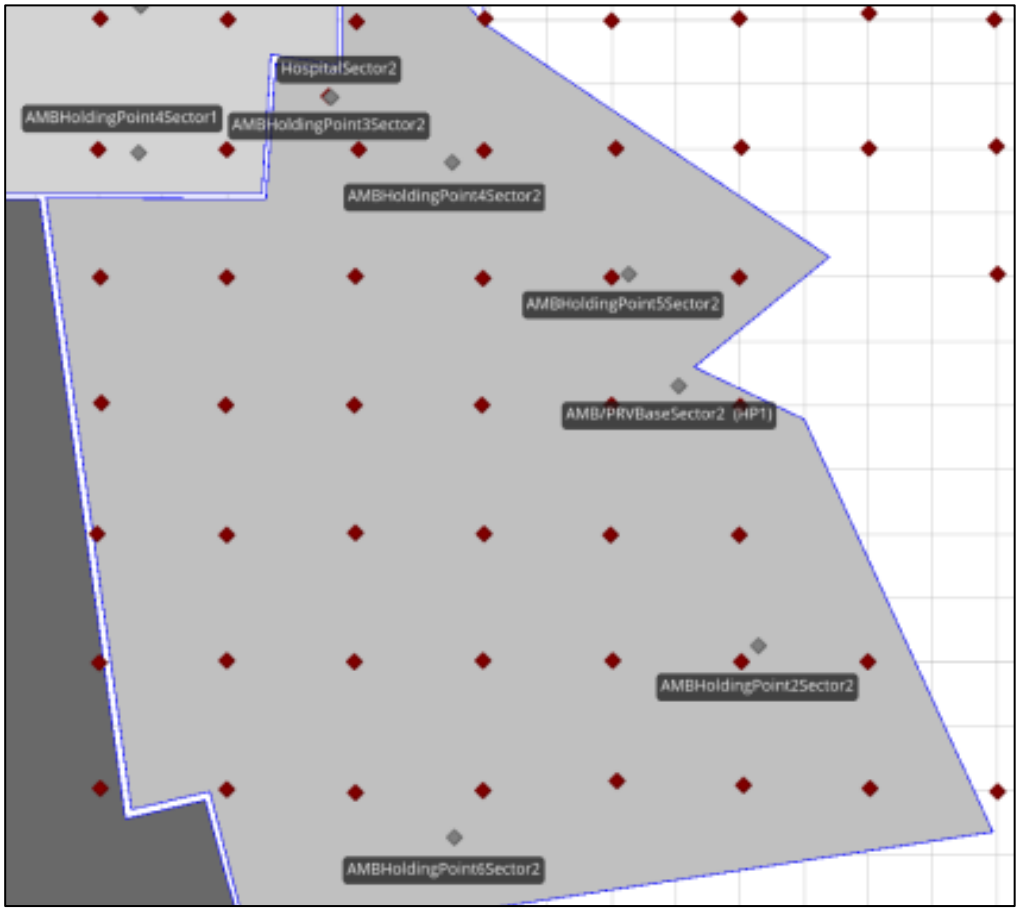


Modelled Representation (Screen Capture: Simio Facility Window)

ANNEXURE I. Sector 2 (GF Jooste Hospital): Map and Modelled Sector

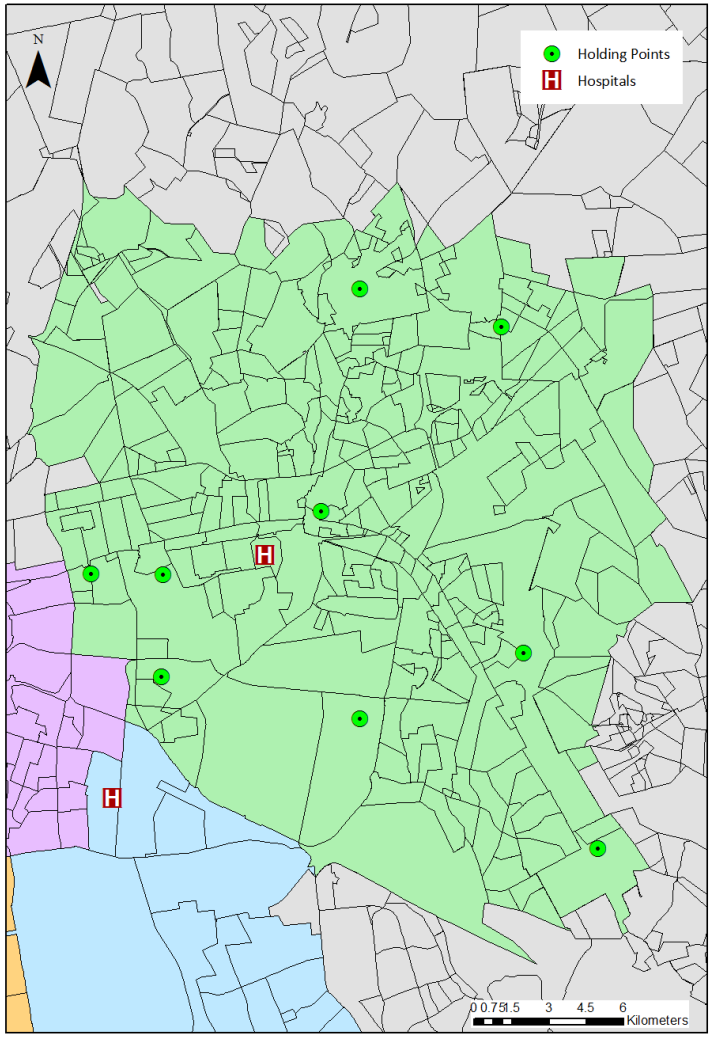


GIS Map

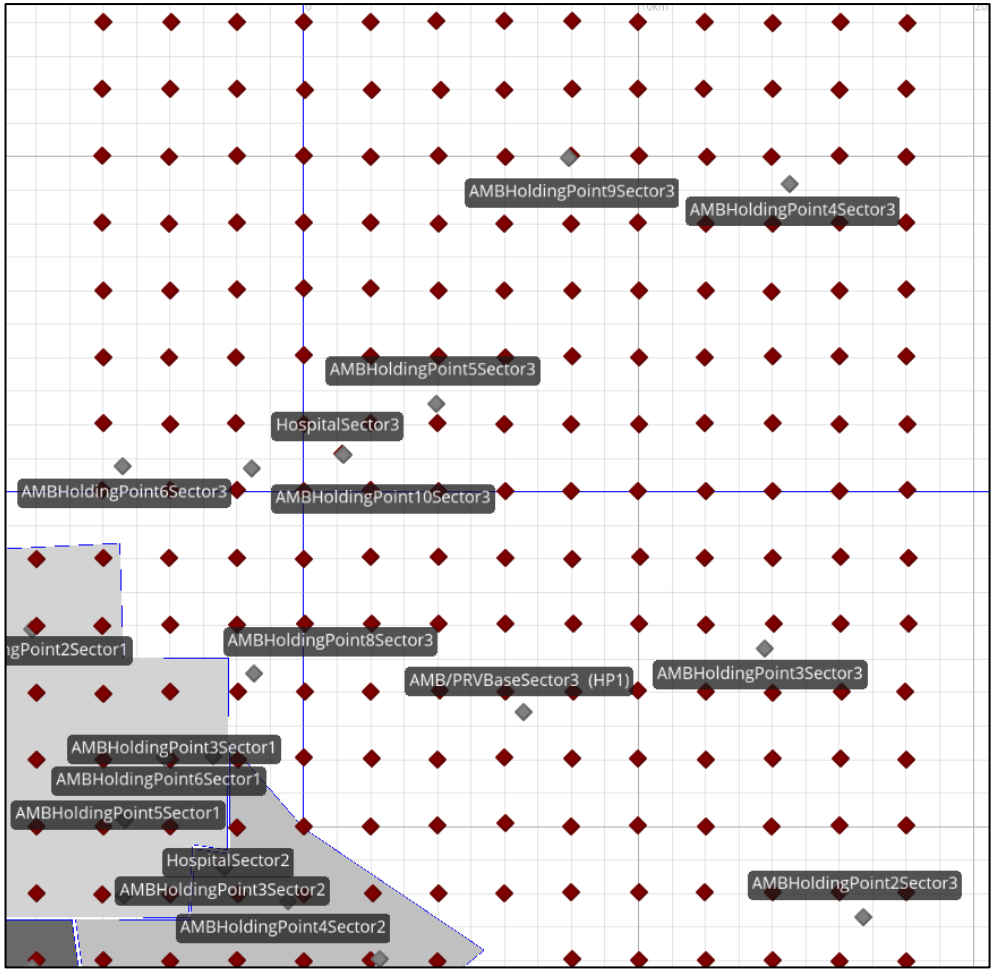


Modelled Representation (Screen Capture: Simio Facility Window)

ANNEXURE J. Sector 3 (Tygerberg Hospital): Map and Modelled Sector

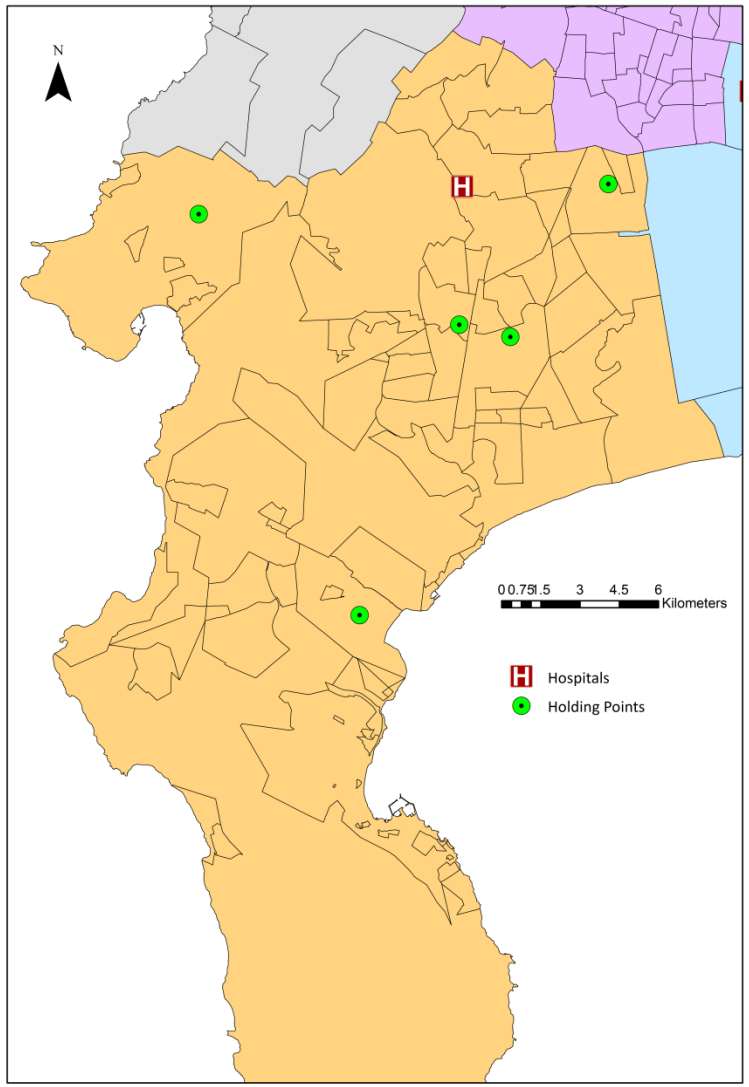


GIS Map

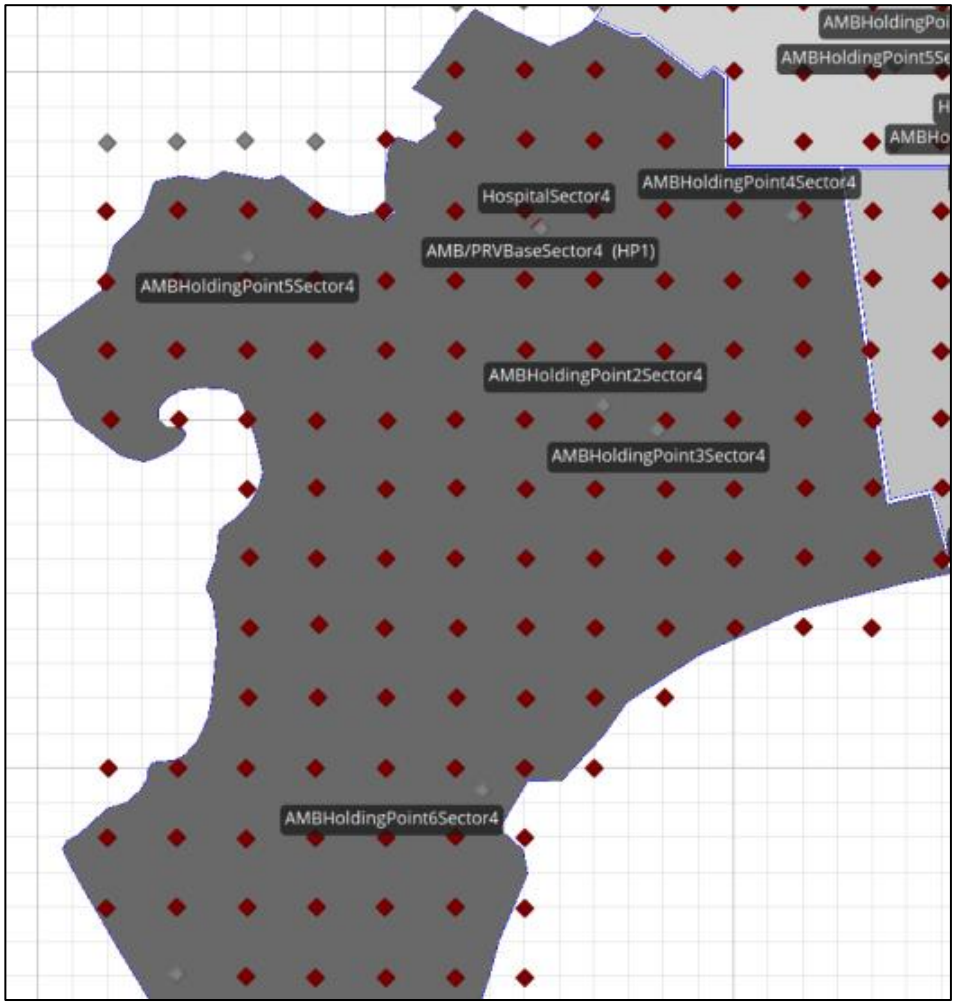


Modelled Representation (Screen Capture: Simio Facility Window)

ANNEXURE K. Sector 4 (Victoria Hospital): Map and Modelled Sector



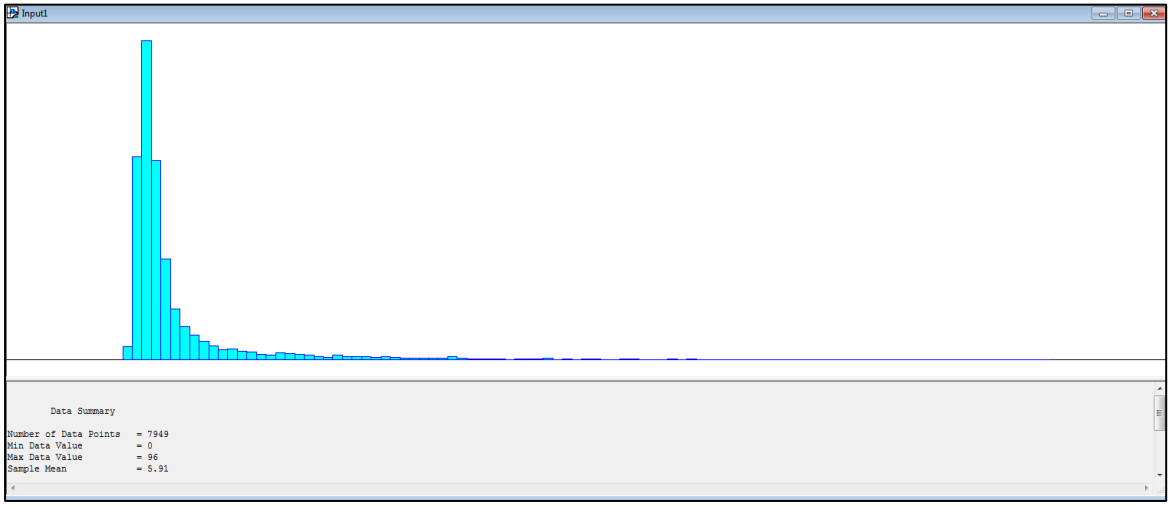
GIS Map



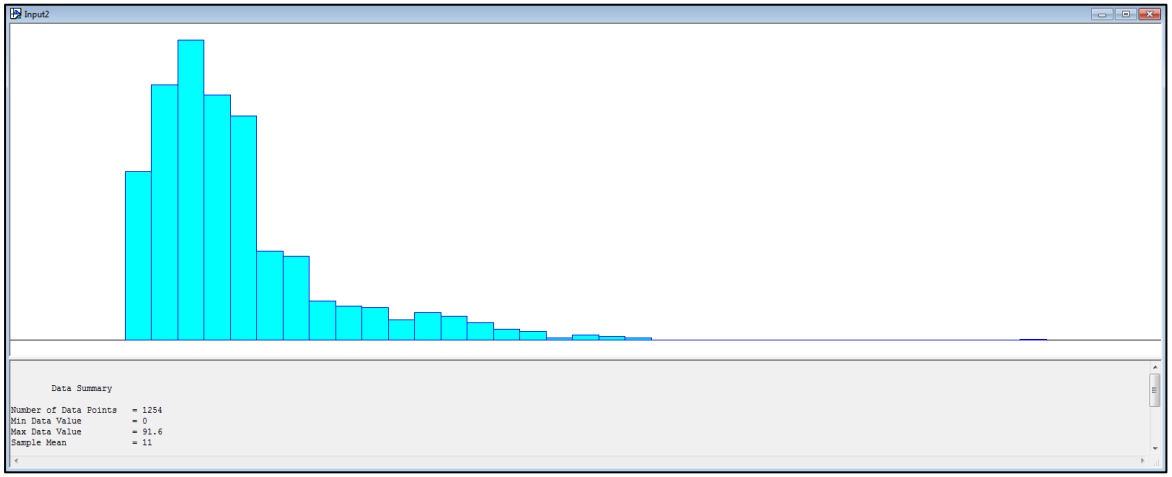
Modelled Representation (Screen Capture: Simio Facility Window)

ANNEXURE L: Sample of Statistical Distributions From Input Data

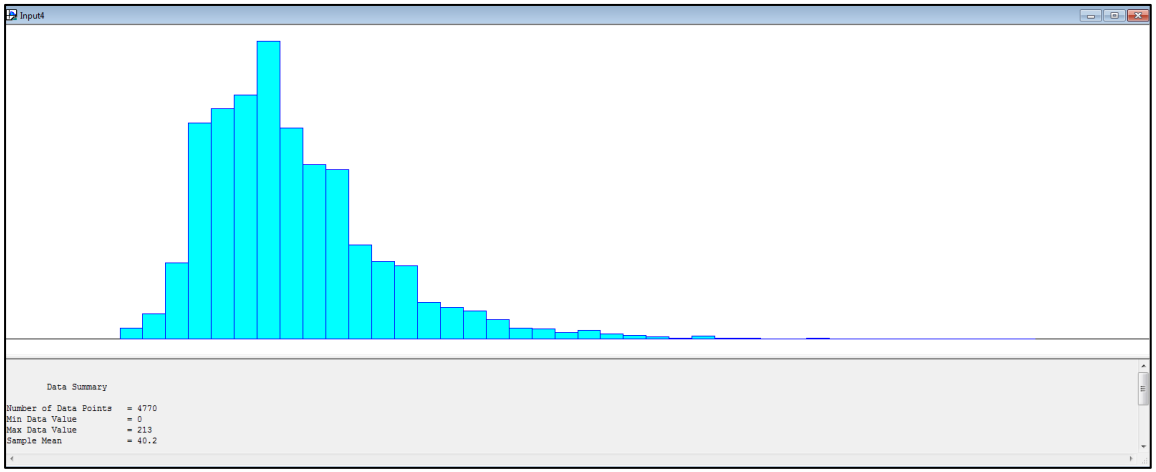
L.1. Priority 1 Dispatch Processing Time



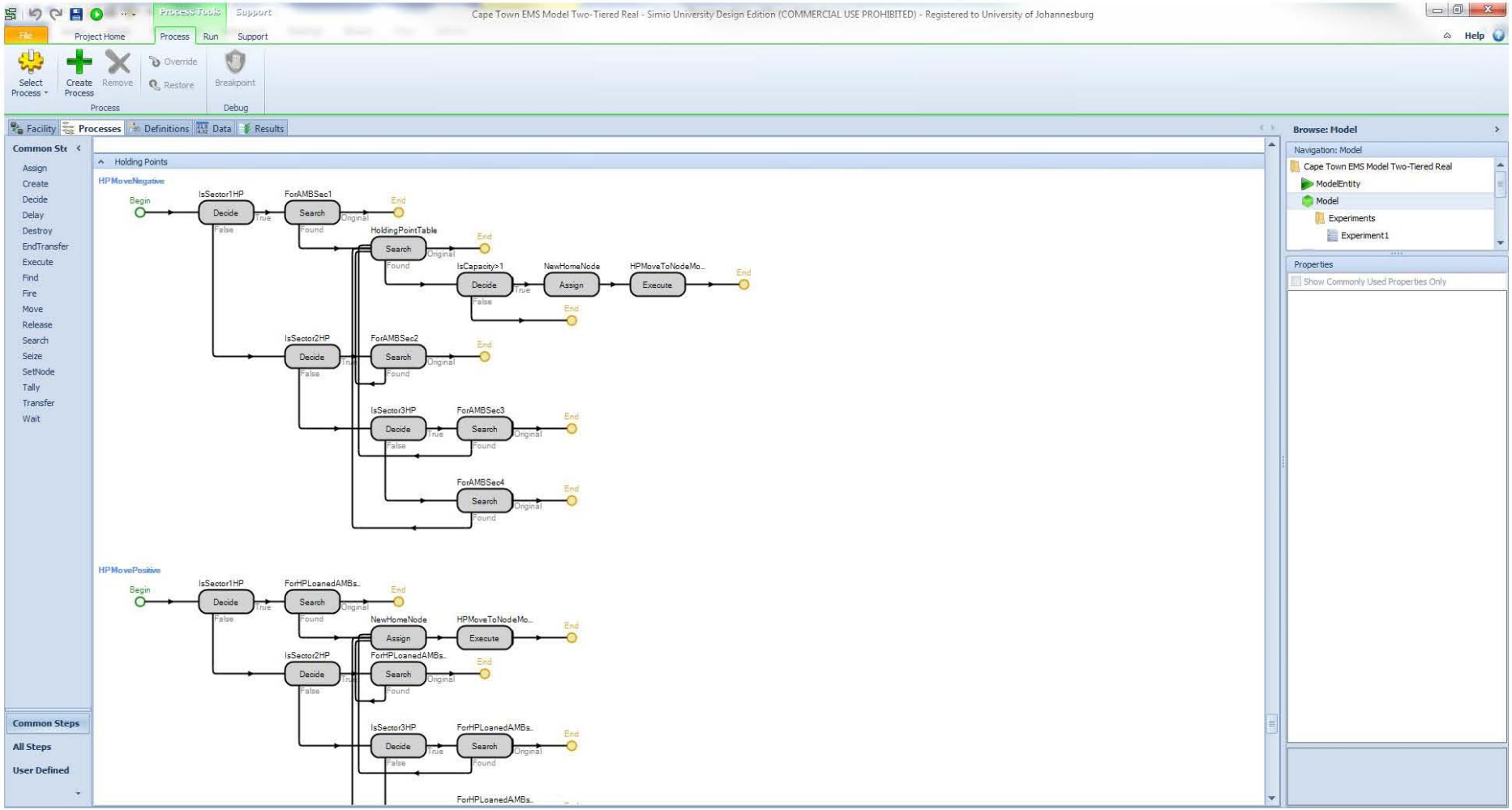
L.2. Priority 1 Response Time



L.3. Priority 1 Scene Time



ANNEXURE M: Example of Simio Process Logic Modelling Environment



ANNEXURE N: Example of Simio Animation